What drives the accuracy of PV output forecasts?

Thi Ngoc Nguyen[1] Autor, Felix Müsgens Co-Autor(en)

Brandenburgische Technische Universität Cottbus - Senftenberg, Adresse: Siemens-Halske-Ring 13, 03046 Cottbus, Telefonnr: +49 (0) 355 69 3227, E-Mail: nguyen@b-tu.de.

Kurzfassung:

This paper provides a statistical analysis on historical Photovoltaic (PV) output forecast performance. A total of 180 papers on PV forecasts have been reviewed for data collection, focusing on the forecast errors, which generates a data base of 1136 observations with 21 key features, covering a variety of models, regions, time sets, level of aggregation etc. This large data base allows harmonising context difference across papers and removing risks of bias from individual studies to come up with a global conclusion regarding "What drives the accuracy of the PV output forecasts". Besides the choice of the methodology, the forecast horizon, the length of the out of sample test sets, the data processing techniques are among the key factors influencing the forecast error levels.

Keywords: PV forecasting, survey paper, inter-model comparison, systematic literature review, statistical analysis

HIGHLIGHTS

- Out-of-sample test set length positively correlates with the forecast errors. An additional day in the test set increases the error by 0.007-0.026 percentage point (pp). The state-of-the-art methods perform more robustly to the change in test set lengths than the classical.
- Long test sets (at least one year) generate more meaningful conclusions on PV output forecast assessment. Restricting the bias from the difference in test set lengths can double the explanation power of the OLS regression from 15% to 35%.
- The possibility of "cherry picking" in reporting errors exists. One-day test sets have the average error value of 2.7% – around a quarter of that of all the other test sets (~10%).
- The longer the forecast horizons are, the higher the forecast errors are. On average, the intra-day and day-ahead forecast errors are higher than the intra-hour by 3.45 pp and 6.12 pp respectively. The state-of-the-art methods have more stable performance when moving from intra-hour to day-ahead forecasts, implying their high potential in improving the long horizon forecasts.
- PV output forecasts have a steady improvement. Models published one year later have the average errors that are 0.64-0.98 pp lower. The progress is more significant for the state-of-the-art than for the classical methods.
- Data processing techniques contributes to enhancing the forecast accuracy. Each one additional technique reduces the average errors by 1.25-1.32 pp. The effect is

[1] Jungautor

stronger for state-of-the-art methods, signalling the further improvement that can be made in the long run by this group of methodologies.

- Among the data processing techniques, data normalization is the most effective, reducing the average error by 3.16 pp, followed by resampling technique (-2.88 pp) and the inclusion of numerical weather predictions (NPW) model's output (-2.48 pp).
- Hybrid, ensemble, and hybrid-ensemble models achieve the lowest forecast errors. Hybrid models are consistently superior to the others and outperform the classical methods by 3.41-3.93 pp. ML performs much worse than expected, having the normalized root mean square error (NRMSE_avg) for day-ahead forecasts increase from 17.5% to 35% when removing the key risks of bias in inter-model comparison.
- The bias caused by context difference can lead to the overestimating of the superiority of the state-of-the-art methods. The complexity-accuracy trade-off therefore favours the classical models in the short and medium run. However, the complex models show much higher potential to enhance forecasts' quality in the long run thanks to the development of new data processing techniques. The future of PV output forecasts is consequently driven by the state-of-the-art models.

List of Abbreviations

Abbreviations	Meaning
2D	2 Dimentional
3D	3 Dimentional
ANFIS	Adaptive Neuro-Fuzzy
ANN	Artifical Neural Network
AR	Auto-Regressive
ARIMA	Auto-Regressive Integrated Moving Average
ARIMAX	Auto-Regressive Integrated Moving Average With Exogeneous Variables
ARMA	Auto-Regressive Moving Average
ARMAX	Auto-Regressive Moving Average With Exogeneous Variables
ARX	Auto-Regressive With Exogeneous Variables
BPNN	Back Propagation Neural Network
CART	Classification And Regression Tree
CFNN	Cascade-Forward Neural Network
CLS	Constrained Least Squares (CLS) Regression
CNN	Convolution Neural Network
CSI	Clear Sky Index
CSLSTM	Clear Sky Index - Long Short Term Memory
CSM	Clear Sky Model
DCNN	Deep Convolution Neural Network
DL	Deep Learning
DNN	Deep Neural Network
ENN	Elman Neural Network
ETS	Exponential Trend Smoothing
FCN	Fully Convolutional Network
FCNN	Fully Connected Neural Network
FFNN	Feed Forward Neural Network
FNN	Feed Forward Neural Network
GA	Genetic Algorithm

GRNN	General Regression Nn
IEA	International Energy Agency
LAD	Least Absolute Deviation
LOESS	Locally-Weighted Regression
LSTM	Long Short Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MARS	Multivariate Adaptive Regression Spline
MBE	Mean Bias Error
ML	Multiple Linear Model
MLP	Multi-Layer Perceptron
MW	Megawatt
NARX	Non-Linear AR-Exogenous
NMAE	Normalized Mean Absolute Error
NRMSE	Normalized Root Mean Square Error
NWP	Numerical Weather Predictions
NZE2050	Net Zero Emissions By 2050
OLS	Ordinary Least Squares
pp	Percentage point
PSO	Particle Swarm Optimization
PV	Photovoltaic
RBFNN	Radial Basis Function Neural Network
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SARIMA	Seasonal Auto-Regressive Integrated Moving Average
SD	Stadard Deviation
SE	Standard Error
SFLA	Shuffled Frog Leaping Algorithm
SVM	Support Vector Machine
TCM	Time Correlation Modification
WT	Wavelet Transform

1 Introduction

We start this part explaining the importance of accurate PV output forecasts to the integration of PV power – the fastest growing energy source on the globe. We then show that there have been so many studies on enhancing the accuracy of PV output forecasts that a survey on the topic is required. Particularly, the question “What drives the accuracy of PV output forecasts?” is crucially important, which has been addressed by most historical reviews on PV output forecasts but remains to be answered concretely. This paper is the first to provide a *statistical analysis* on PV output forecasts’ errors to answer this question.

Electricity generation from renewable energies is projected to overtake coal by 2025 and to provide up to 80% of the global electricity demand growth to 2030 (IEA, 2020). The use of renewable energies as the main supply of power is decisive to reaching the target of net zero emissions globally by 2050. Among many forms of renewable energies, solar energy is the most important and has become systematically cheaper than all other power sources in most countries (IEA, 2020). Estimates from the World Energy Outlook 2020 show an

increase of 13% per year in photovoltaic (PV) sector, which can supply one-third of the growth in the global electricity demand from 2020 to 2030. Consequently, the International Energy Agency refers to PV as the “new king” of electricity generation.

Electricity generation from PV, which is often referred to as “PV output” in the literature, comes from the sun, using PV modules to convert solar irradiance to electricity. For this reason, PV output depends largely on solar irradiance and is vulnerable to the change in meteorological variables such as temperature, wind speed, cloud cover, atmospheric aerosol levels and humidity, which are by nature particularly stochastic (Raza *et al.*, 2016). This leads to a high volatility in PV output, causing difficulty in managing and planning power plant operations and blocking new investment for fear of system instability. To integrate PV output into the global power supply, this variability and uncertainty must be dealt with.

Having high quality PV output forecasts has emerged as a particularly efficient solution for this problem (Ahmed *et al.*, 2020; Das *et al.*, 2018; Pazikadin *et al.*, 2020; Raza *et al.*, 2016). The better PV forecasts are, the more PV can be integrated into the system and the better can power plant operations be planned, saving money e.g., on start-up costs, and the higher the reliability of grid operation (i.e., lower risk of network failure or less balancing power needed). Consequently, millions of USD per year are spent on forecasts, software tools and methods, and on their improvement. At the forefront of this commercial applications, academic researchers have published hundreds of papers on these issues, enabling further progress in this field. While this re-confirms the importance of the topic, it also makes keeping track of most important new advances and comparing the newest forecasts techniques to the existing work more challenging.

This leads to a demand for systemizing the scientific knowledge with regard to: What drives the accuracy of PV output forecasts? A concrete and global answer to this question is important as it allows scholars to adapt their research agenda, and both investors and system planners to know more about the respective forecast error to expect.

Indeed, 12 out of a total of 13 review papers on PV output forecasts that we could find[2] (Table 1) have aimed at answering the above question through summarising the findings from individual studies, discussing the pros and cons of different methodologies, or comparing models’ performance. However, these reviews have not considered the risks of bias that can be caused by the context difference among papers such as the error calculation formula, the length of the test sets…. This review approach therefore potentially contains bias and produces more qualitative rather than concrete conclusions.

In the meanwhile, the risks of bias caused by individual studies in PV output forecast are high because the PV output forecast performance depends strongly on contextual factors ((Ahmed *et al.*, 2020; Blaga *et al.*, 2019) and the forecasters rarely use identical data sets or standardised error report methods (Ahmed *et al.*, 2020; Antonanzas *et al.*, 2016; Yang *et al.*, 2018). Consequently, any conclusion made from reviewing individual papers necessarily requires harmonising context distinctions and removing risks of bias, which can be achieved through a transparent process of collecting relevant research, screening the quality of

[2] Papers collected from Google Scholar using keyword “Review on PV output forecasts”.

papers, extracting the secondary data base, and analysing the (possibly unbiased) data using statistical methods. We call this process a “statistical analysis”.

Using statistical analysis to integrate findings from individual studies has enjoyed a surge in popularity in many disciplines such as clinical medicine, social policy, education, information systems, and software engineering (Brereton *et al.*, 2007; Chavez Velasco *et al.*, 2021) and has been suggested as a crucially important body of research as it allows researchers and decision-makers to rigorously synthesize the outcomes from historical studies in an objective and evidence-based manner (Borenstein *et al.*, 2009).

Surprisingly, there has been no statistical analysis for PV output forecasts historically. In fact, to allow this statistical analysis approach to generate concrete and global conclusions requires that the number of papers available for reviewing be so large that there are sufficient observations in different contexts for analysis. Otherwise, reviewers may end up having too few observations in each context to draw any conclusion. Fortunately, with the importance of the PV output forecasts, so much effort and progress has been made in this field so far that it now enables and requires a statistical analysis. This motivates us to start this paper and to give a concrete and global answer to the question of what factors can enhance the PV output forecasts’ accuracy.

This paper contributes in the following ways:

First, this paper is the first to build a data base of PV output forecasts’ errors on a large scale of 180 papers from 2007 until now using a well-defined and transparent process of collecting and screening the quality of relevant research. All the papers are read and only papers qualifying certain requirements such as sufficient information or appropriate approach of forecasting[3] are included for data extraction and analysis. After processing the data, we have a data base of 1136 observations, with each observation being the average error reported by a specific model in a paper, featured by 21 variables including the time and place of the data sets, the forecast horizon and resolution, the error metrics and normalization methods, and other variables related to methodologies such as the type of model and data processing techniques…

Second, this is also the first paper to provide a statistical analysis on PV output forecasts to identify the factors that can enhance the forecast quality and to compare the inter-model performance. Particularly, this is the first attempt to examine the claims made in historical literature reviews on PV output forecasts that remain to be confirmed using statistical analysis. In other words, this is the first “survey of surveys” on PV output forecasts.

Finally, through the statistical analysis, the paper provides many findings that are important to the further progress of PV output forecasts:

- Out-of-sample test set length positively correlates with the forecast errors.
- Long test sets (at least one year) generate more meaningful conclusions on PV output forecast assessment.
- The possibility of “cherry picking” in reporting errors exists.

[3] Details are explained in section 3.1.2.

- The longer the forecast horizons are, the more difficult to have high forecast accuracy.
- PV output forecasts have a steady improvement in accuracy throughout the time.
- Data processing techniques contributes to enhancing the forecast accuracy, with the best candidates being the technique of data normalization, resampling, and the inclusion of NWP model's output.
- Hybrid, ensemble, and hybrid-ensemble models achieve the lowest forecast errors while ML performs much worse when removing the key risks of bias in inter-model comparison.
- The superiority of the state-of-the-art methods can be overestimated if we do not consider the risks of bias caused by context difference. The complexity-accuracy trade-off therefore favours the classical models in the short and medium run. However, the complex models show much higher potential to enhance forecasts' quality in the long run thanks to the development of new data processing techniques. The state-of-the-art are also more robust to the change in test set lengths and forecast horizons. The future of PV output forecasts is consequently driven by the state-of-the-art models.

Besides, this paper indicates that carrying out this statistical analysis is costly, though necessary, and emphasizes that establishing a benchmark for assessing PV output forecast performance is the next step to save time and resources for future knowledge systemization.

The structure of this paper is as follow: Section 2 discusses the historical reviews on PV output forecasts and other related works. Section 3 describes the methodology and the data base. Section 4 presents the data analysis and provides important implications. Section 5 briefly discusses the benchmark for PV output forecast assessment, and section 6 concludes the paper.

2 Literature review

In this part, we first briefly present the historical review works on PV output forecasts that we found and summarise scholars' opinions on what factors drive the PV output forecasts' accuracy. These opinions will be examined through the statistical analysis in the section 4. Then we discuss one specific work that was not on PV output forecasts but has an approach that is close to ours and how we depart from there.

2.1 Historical review papers on PV output forecasts

Using Google Scholar with the keyword "Review on PV output forecasts", we found a total of 13 review papers on PV output forecasts. Table 1 summarises the key points of these review papers.

Table 1: Historical reviews on PV output forecasts

No	Authors (Year)	Summary
1	Ahmed *et al.* (2020)	A review on the short-term PV output forecasts and highly advanced methodologies such as hybrid models using the latest techniques. It suggests that factors such as time stamp and forecast horizon, and techniques of data processing, weather classification, and parameter optimization can influence the quality of the forecasts and should be taken into account when comparing models.
2	El hendouzi and Bourouhou (2020)	A review on short-term PV output forecasts that discusses the basic principles, standards, and different methodologies of PV output forecasting.
3	Mellit *et al.* (2020)	A complete and critical review on highly advanced methods for PV output forecasts, especially the recent development in machine learning (ML), deep learning (DL), and hybrid methods.
4	Pazikadin *et al.* (2020)	A review of 87 articles on both solar irradiance and PV output forecasts, focusing on artificial neural network (ANN)-based models only. It highlights the superiority of the ANN hybrid models and emphasizes the importance of data input quality and weather classification.
5	Rajagukguk *et al.* (2020)	A review of DL models for PV output forecasts and solar irradiance forecasts. It compares 3 individual deep learning models and one hybrid model using deep learning techniques and shows that the hybrid outperforms the 3 individuals. It also recommends papers using normalized errors to enable inter-model comparison.
6	Akhter *et al.* (2019)	A review on ML and hybrid methods for solar irradiance and PV output forecasts that suggests the superiority of ML-based hybrid models.
7	Das *et al.* (2018)	A review on the development in PV output forecasts and model optimization techniques. It suggests that ANN and support vector machine (SVM)-based models have very good and robust performance.
8	Sobri *et al.* (2018)	A review on PV output forecast methods that indicates the superiority of ANN and SVM-based models. It also suggests ensemble methods have much potential in enhancing the forecast accuracy.
9	Yang *et al.* (2018)	A review on both solar irradiance and PV output forecasts using text mining, focusing on the analysis of the features of models and predicting the trend in PV forecasting.
10	Barbieri *et al.* (2017)	A review on very short-term PV output forecasts with cloud modelling. It suggests that hybrid models combining physical with statistical models can enhance the forecast accuracy, especially when PV outputs have rapid fluctuations.
11	Antonanzas *et al.* (2016)	A review on PV output forecasts that suggests the dominance of ML-based models.
12	Raza *et al.* (2016)	A discussion of ML-based and classical methods for PV output forecasts that supports the use of ML models and data processing techniques.
13	Mellit and Kalogirou (2008)	The first review on ANN-based models for PV output forecasts that suggests a high potential of ML techniques in enhancing the forecast accuracy.

Most of these review papers address the question of what factors driving the PV output forecast accuracy and agree that both the methodology and the empirical set-up influence the forecast accuracy (e.g., Ahmed *et al.* (2020), Pazikadin *et al.* (2020), Raza *et al.* (2016), Das *et al.* (2018), Mellit *et al.* (2020)). Following we summarise the key arguments suggested by scholars regarding the factors influencing the PV output forecast accuracy.

First, the PV output forecast accuracy is negatively correlated with the length of the test set (Ahmed *et al.*, 2020). A shorter test set usually means less fluctuation in weather conditions and thus higher forecast accuracy (e.g., forecasts made for one single season can be more accurate than made for the whole year of 4 different seasons). Furthermore, reporting errors on a small number of days possibly makes “cherry picking” easier, i.e., to focus on specific days when models achieve the lowest errors. Therefore, many scholars recommend that the test set be at least one year so that the models can show a robust and unbiased performance (Raza *et al.*, 2016).

Second, forecast horizons are negatively correlated with the forecast accuracy (Raza *et al.*, 2016). The forecast horizon measures the time that the forecast looks ahead (Das *et al.*,

2018), which lies between the moment the forecast is made and the moment that the forecast is meant for[4]. Because of the stochastic nature of the meteorological variables that strongly influence the PV output, the longer the forecast horizons are, the more difficult for the forecasts to be precise. Forecast errors are therefore expected to increase with the forecast horizon length (Ahmed *et al.*, 2020; Akhter *et al.*, 2019).

Third, all scholars also agree that huge progress has been made in reducing the PV output forecasts errors during the last decade (Ahmed *et al.*, 2020; Blaga *et al.*, 2019; Mellit *et al.*, 2020; Raza *et al.*, 2016). This corresponds well to the increasing importance of PV energy in the global power supply. Therefore, the literature claims that the accuracy of PV output forecasts is positively correlated with time, that is, the later a paper is published, the lower the forecast errors (on average) should be.

Fourth, in addition to the above factors, data processing techniques[5] can significantly improve the quality of the forecasts (Ahmed *et al.*, 2020; Akhter *et al.*, 2019; Mellit *et al.*, 2020; Pazikadin *et al.*, 2020; Raza *et al.*, 2016). Particularly, cluster-based algorithms, wavelet transform (WT), and the use of numerical weather prediction (NWP) variables are assessed as the most efficient data processing techniques (Ahmed *et al.*, 2020).

Finally, comparing the performance of different methodologies[6], scholars have been particularly optimistic about state-of-the-art methods such as ML and hybrid models in improving the PV output forecasts, agreeing that these models can utilize the advantages of both linear and non-linear techniques and therefore can achieve the best performance for all forecast horizons (Ahmed *et al.*, 2020; Das *et al.*, 2018; Leva *et al.*, 2017; Pazikadin *et al.*, 2020; Rajagukguk *et al.*, 2020; Raza *et al.*, 2016). Therefore, despite the higher complexity and computational burden, many suggest that it is worth investing in complex models in the long run (Ahmed *et al.*, 2020; Akhter *et al.*, 2019).

Interestingly, no solid conclusions have been made regarding these above opinions. Through the statistical analysis of the data of PV output forecast errors, we will examine these claims and quantify the effects of different factors on PV output forecast accuracy.

---

[4] So far there has been no official classification of forecast horizons (Raza *et al.* (2016); Sobri *et al.* (2018)). However, there are two key approaches of horizon classification according to Ahmed *et al.* (2020): (1) very short-term or ultra-short term (from seconds to less than 30 min), short-term (30 minutes to 6 hours), medium-term (6 to 24 hours) and long-term (>24 hours); and (2) intra-hour or nowcasting (a few seconds to an hour), intra-day (1 to 6 hours) and day ahead (>6 hours to several days). The second approach is specifically for PV output forecasts and this paper follows this classification.

Note that forecast horizon is different from forecast resolution. Forecast resolution measures the amount of time between the individual forecasts within one horizon. For example, a forecast of day-ahead horizon and 1 hour resolution is the forecast that predicts the value for every hour the next day.

[5] The description of all data processing techniques that we observe from reviewed papers is presented in Table 2.

[6] How models are classified in this paper is presented in Appendix B.

2.2 Depart from a statistical analysis on solar irradiance forecasts

To the best of our knowledge, Blaga *et al.* (2019)'s work is the closest to our approach, who focused exclusively on solar irradiance forecasts rather than PV output[7]. Blaga *et al.* collects data from 40 papers between 2007 and 2016 and analyses the performance of models in different forecast horizons using two key error metrics namely root mean square error (RMSE) and mean bias error (MBE) both normalized by average values of solar irradiance. At least from the perspective of our review on PV output forecasts, we see that there are some important points to be improved from Blaga *et al.*'s work that apply to any other statistical analysis on energy forecasting as follows:

First, distinguishing clearly between daily and hourly forecasts is important as the former are much easier and combining them may thus distort results. Second, scholars should refrain from filling missing information, particularly the important features such as the error normalization method. Changing from one normalization method (e.g., installed capacity) to the other (e.g., average power value) can greatly change the level of the errors, therefore any assumption on such information can bias the results. Third and the most importantly, more factors should be taken into account when comparing the errors of models. For example, though many scholars suggest the correlation of forecast errors and the test set length, no review work in the past (including Blaga *et al.*'s work) has considered this factor when comparing models' performance.

In section 3, we illustrate how these improvements are implemented in our paper.

3 Methodology and data

In this part, we first illustrate the process of conducting the statistical analysis on PV output forecasts and then we give an overview of the data base that we extracted from the reviewed literature.

3.1 Conducting the statistical analysis on PV output forecasts

We conduct the statistical analysis in four steps. In the first place, we identify and collect the relevant research using Google Scholar. Then we carry out a preliminary examination on the quality of all the papers and remove or include papers for review based on well-defined requirements that will be explained below. Next, we extract the data and have processing

[7] To forecast PV power includes two basic approaches: the direct and indirect one. The direct approach forecasts PV power directly from historical data of PV power and is usually combined with meteorological parameters. The indirect approach first forecasts solar irradiance and then calculate PV power from the solar irradiance using specialized software such as TRNSYSM, PVFORM, and HOMER (Dalton *et al.* (2009). Solar irradiance forecasts are thus an important part of indirect PV forecasts and account for a large proportion of studies on PV power forecasts. Therefore, many scholars include both irradiance forecasts and PV output (electricity) forecasts in their review on PV power forecasts (e.g., Antonanzas *et al.* (2016); Yang *et al.* (2018); Pazikadin *et al.* (2020)). However, the forecasting of solar irradiance is only a step in the whole process of PV output indirect forecasting and therefore should not be identified as PV output forecasting.

steps as necessary. Finally, we analyse the data base using OLS regression and other data visualization techniques to answer the research question “What drives the accuracy of the PV output forecasts?”. The whole process is illustrated in the Figure 1.

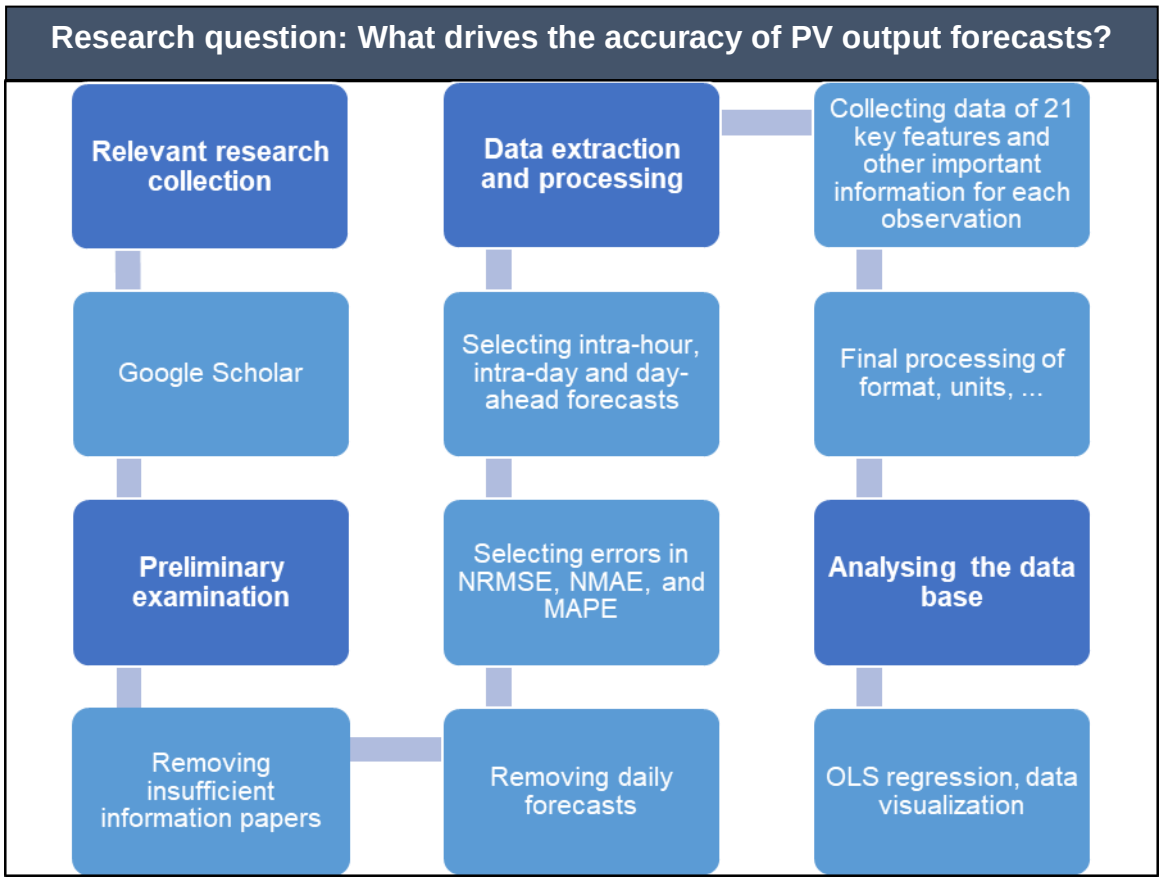

Figure 1: Conducting the statistical analysis on PV output forecasts

3.1.1 Relevant research collection

As discussed in the introduction, we focus our review on the PV output forecasts and would like to examine what factors are important to enhancing the forecast’s accuracy. To search for all available papers on PV output forecasts, we use Google Scholar with the keywords as follows:

- General PV forecasts: combinations of [“PV forecast” or “PV output forecast” or “PV output prediction” or “PV power generation forecast” or “PV power generation prediction”] and [“intra-hour” or “hour-ahead” or “intra-day” or “day-ahead”].
- Classical methodologies: combinations of [keywords for General PV forecasts] and [“regression” or “autoregression” or “removing seasonality” or “detrending” or “AR(X)” or “ARMA(X)” or “ARIMA(X)” or “bootstrap”].
- ML: combinations of [keywords for General PV forecasts] and [“machine learning” or “artificial neural networks” or “ANN”].
- Hybrid methods: combinations of [keywords for General PV forecasts] and [“hybrid” or “ensemble” or “advanced”].

Among the search results, we could collect a total of 180 papers on PV output forecasting published from 2007 until 2020.

3.1.2 Preliminary examination

In the second step, we read all 180 papers and conduct the quality check as follows:

(i) Remove papers of insufficient information

In the first place, we observe that there are many papers not providing sufficient or clear information. For example, some papers do not mention whether they are doing daily or hourly forecasts, and some others do not give information on forecast horizons. There are also many papers unclear about their calculation of the forecast errors.

We therefore require the key information to be provided in the papers. Particularly, papers have to report the errors for PV output forecasts. Furthermore, the information of the forecast horizon, forecast resolution, the test set, and other information for inter-model comparison such as the error normalization and calculation method must be clearly explained. All the papers that do not provide sufficient information as required are excluded.

(ii) Remove papers providing daily forecasts

As discussed above, hourly forecasting should be analysed separately from daily forecasting. We also observe from the 180 papers that most of the studies focus on hourly or less than 1 hour resolution forecasts. In this paper, we therefore aim at examining the models that can provide no longer than 1 hour resolution forecasts and exclude all the others, including the daily forecasts.

(iii) Keep only papers reporting (or providing information to calculate) normalized RMSE, mean absolute error (MAE) and mean absolute percentage error (MAPE)

We found at least 18 types of metrics that have been used by scholars to measure the performance of a PV output forecast model. Among these, RMSE, MAE, and MAPE are the mostly chosen. To enable error metric comparison across models, the first condition is to have the same error metric, and the second condition is to have the error metrics normalized. Therefore, we took only the papers that report at least one of these 3 error metrics in normalized values, or in case they provide absolute error values, additional information to calculate normalized errors must be provided (e.g., installed capacity or peak power of the plant).

Besides, we also observe that scholars can have different calculation approaches for a same error metric, especially regarding the normalization method of the errors. For example, the RMSE can usually be normalized by the measured value of the data, the average value, the maximum measured value, the installed capacity of the plant, or simply by normalizing the data set and calculating the errors from the normalized data. In many cases, scholars simply report normalized errors without defining their calculation mechanism or the reference quantity for error normalization. All the papers that report the error metrics but use the calculation approach different from the standardised[8] or not clearly explain their calculation formula are also excluded.

(iv) Keep only intra-hour, intra-day and day-ahead horizons[9]

[8] Details on the standardised formulas of error metrics are presented in Appendix C.

[9] See Table 2 for horizon classification

Examining the papers, we observe that the number of forecasts longer than two days ahead is too low to be included in the data base. Therefore, we keep only the papers providing forecasts for intra-hour, intra-day and day-ahead horizons.

At the end of the preliminary examination, we have 66 papers left for data extraction[10].

3.1.3 Data extraction and processing

The 66 papers are examined thoroughly for data extraction. For each observation that is the average error reported by a certain model in a paper, we collect the information for at least 21 different features, which are summarised in Table 2.

Then we carry out data processing steps such as harmonising the units (e.g., W, kW, MW), normalizing errors based on available information, and fixing the data format. At the end of this process, a data base of 1,136 observations is built for further analysis.

3.1.4 Data analysis

We first examine the effects of all factors of interest on the PV output forecasts' accuracy by doing OLS regressions with the dependent variable being the average error of PV output forecast models (E) and the explanatory variables include the test set length (TL), the three dummy variables for forecast horizon including intra-hour, intra-day and day-ahead horizons (H), the publishing year of the paper (Y), the complexity of the model (C), the seven dummies of the type of the models (M), and the eleven dummies of data processing techniques (T). These explanatory variables are the key factors that are suggested by many scholars to influence the forecast accuracy, as discussed in section 2.1. By quantifying the effects of these key factors through the OLS regression, we can systematically survey the historical surveys on PV output forecasts. The regressions are represented by the following two equations:

$$E = \beta_0 + \beta_1 TL + \sum_{i=1}^{3} \beta_{i+1} H_i + \beta_5 Y + \beta_6 C + \sum_{j=1}^{7} \beta_{j+6} M_j + \varepsilon \quad (1)$$

$$E = \beta_0 + \beta_1 TL + \sum_{i=1}^{3} \beta_{i+1} H_i + \beta_5 Y + \sum_{j=1}^{11} \beta_{j+5} T_j + \sum_{k=1}^{7} \beta_{k+16} M_k + \varepsilon \quad (2)$$

Equation (1) describes the main OLS regression that goes along the whole analysis of all factors of interest, with the left-hand side representing the dependent variable and the right-hand side including the explanatory variables and the error term (ε). β is the coefficient of the explanatory variable, which will be computed through the OLS regression and informs the effect of each explanatory variable on the forecast errors.

This main regression is done first on the whole data base (regardless of the error metric) and then on this same whole data base but restricting only observations of test sets at least one year. Besides investigating the impacts of varied factors on error values, this is also to examine the importance of having a long test set as claimed by many scholars. Showing that the length of the test sets has an influence on the forecast error levels is particularly important, as this is the most obvious indication of the bias that exists in generalizing

[10] The list of 66 papers is presented in Appendix A

conclusions from individual papers without harmonising the context difference as discussed at the very beginning of this paper. The threshold of one year is used as it is suggested by many scholars to sufficiently test models' robust performance (Raza *et al.*, 2016). Then we also conduct regressions on subsets of classical and state-of-the-art models to examine the difference in the effects of the factors on different groups of methodologies, which will provide important implications for later inter-methodology comparison.

Equation (2) describes a modified version of the main regression, which focuses on quantifying the effects of individual data processing techniques (rather than the number of techniques used) on the forecast accuracy. In this modified regression, the variable of complexity is therefore replaced by the variables of data processing techniques. The results of this regression reveal which technique is more efficient and contribute significantly to further improvement of PV output forecasts.

For each (explanatory) variable, we also use boxplot and other data visualization methods to visualize its effects on PV output forecasts for different subsets of data, where the difference in context and thus the risks of bias are eliminated. This not only helps picture the conclusions, but it also makes the conclusions solid.

3.2 Data description

This part provides a brief description on our data base. Each data point is featured by 21 statistical variables including the information of the publishing year of the papers where the data point is collected (Var. 1), the error values (Var. 2), 10 data processing techniques (Vars. 3-13), the length of the test sets (Var. 14), the forecast resolution (Var. 15), and the complexity of the model (Var. 16) and 5 categorical variables (Vars. 17-21) namely Country, Region, Methodology, Forecast Horizon, and Error Metric as described in the Table 2.

Table 2: Data description

Statistical Variables													
No	**Vars**	**Unit**	**Description**	**Obs.**	**Mean**	**SD**	**Median**	**Min**	**Max**	**Range**	**Skew**	**Kurtosis**	**SE**
1	Publishing Year	NA	The year that the paper is published	1136	NA	NA	2019	2007	2020	13.00	-1.30	0.47	0.08
2	Error	%	The average error reported for the model in the paper	1136	9.19	9.77	8.03	0.00	100.47	100.47	3.38	20.24	0.29
3	Transformation	Times used (1 if the technique is used in the model and 0 otherwise)	Use WT or any other techniques to transform or decompose data to remove spikes or high fluctuation in the data	1136	0.08	0.28	0.00	0.00	1.00	1.00	3.00	7.03	0.01
4	Normalization		Bring variables of varied ranges and units to the same range of [-1,1] or [0,1] without unit for easy comparison and modelling	1136	0.46	0.50	0.00	0.00	1.00	1.00	0.18	-1.97	0.01
5	Outlier		Use techniques to handle outliers	1136	0.01	0.07	0.00	0.00	1.00	1.00	13.63	184.01	0.00
6	Cluster-based		Use cluster-based techniques such as k-means to pre-process data	1136	0.35	0.48	0.00	0.00	1.00	1.00	0.61	-1.63	0.01
7	NWP-related		Include NWP variables among inputs or use NWP to classify weather conditions before forecasting	1136	0.37	0.48	0.00	0.00	1.00	1.00	0.54	-1.71	0.01
8	CSI		Use CSI in data pre-processing	1136	0.13	0.34	0.00	0.00	1.00	1.00	2.17	2.72	0.01
9	Spatial average		Data pre-processing techniques to reduce fluctuations of forecast	1136	0.01	0.09	0.00	0.00	1.00	1.00	10.50	108.41	0.00
10	Resampling		Resample the data to diverse the training sets	1136	0.13	0.33	0.00	0.00	1.00	1.00	2.22	2.92	0.01
11	Weather forecast		Use weather forecast to classify weather before forecasting	1136	0.00	0.06	0.00	0.00	1.00	1.00	16.74	278.51	0.00
12	Regression		Use regression to analyse data before forecasting	1136	0.00	0.03	0.00	0.00	1.00	1.00	33.62	1129.01	0.00
13	Dimension reconstruction		Reconstruct dimensions of data (e.g., 2D to 3D)	1136	0.01	0.07	0.00	0.00	1.00	1.00	13.63	184.01	0.00
14	Test set length	Days	The length of the data set used for testing the model and calculating the	1136	214.45	235.60	90.00	1.00	730.00	729.00	1.11	0.05	6.99

			error										
15	Resolution	Minutes	The time interval between the individual forecasts within one horizon	1136	43.42	24.04	60.00	1.00	60.00	59.00	-0.79	-1.32	0.72
16	Complexity	NA	Count the number of data processing techniques used in the model	1136	1.55	1.33	1	0	4	4	0.64	-0.86	0.04

Categorical Variables

No	**Vars**	**Description**
17	Country	The country of the data set used for training and testing the model
18	Region	The region of the data set used for training and testing the model
19	Methodology	The classification of models (Appendix B for more detail)
20	Forecast Horizon	The time that the forecast looks ahead, i.e., between when the forecast is made and when the forecast is meant for. This paper classifies horizons into intra-hour or nowcasting (a few seconds to an hour), intra-day (1 to 6 hours) and day ahead (>6 hours to several days).
21	Error metric	The error metric reported by the paper, including the normalization methods (Average or measured values (_avg), Installed capacity or peak power (_installed), and Normalized Data (_norm)).

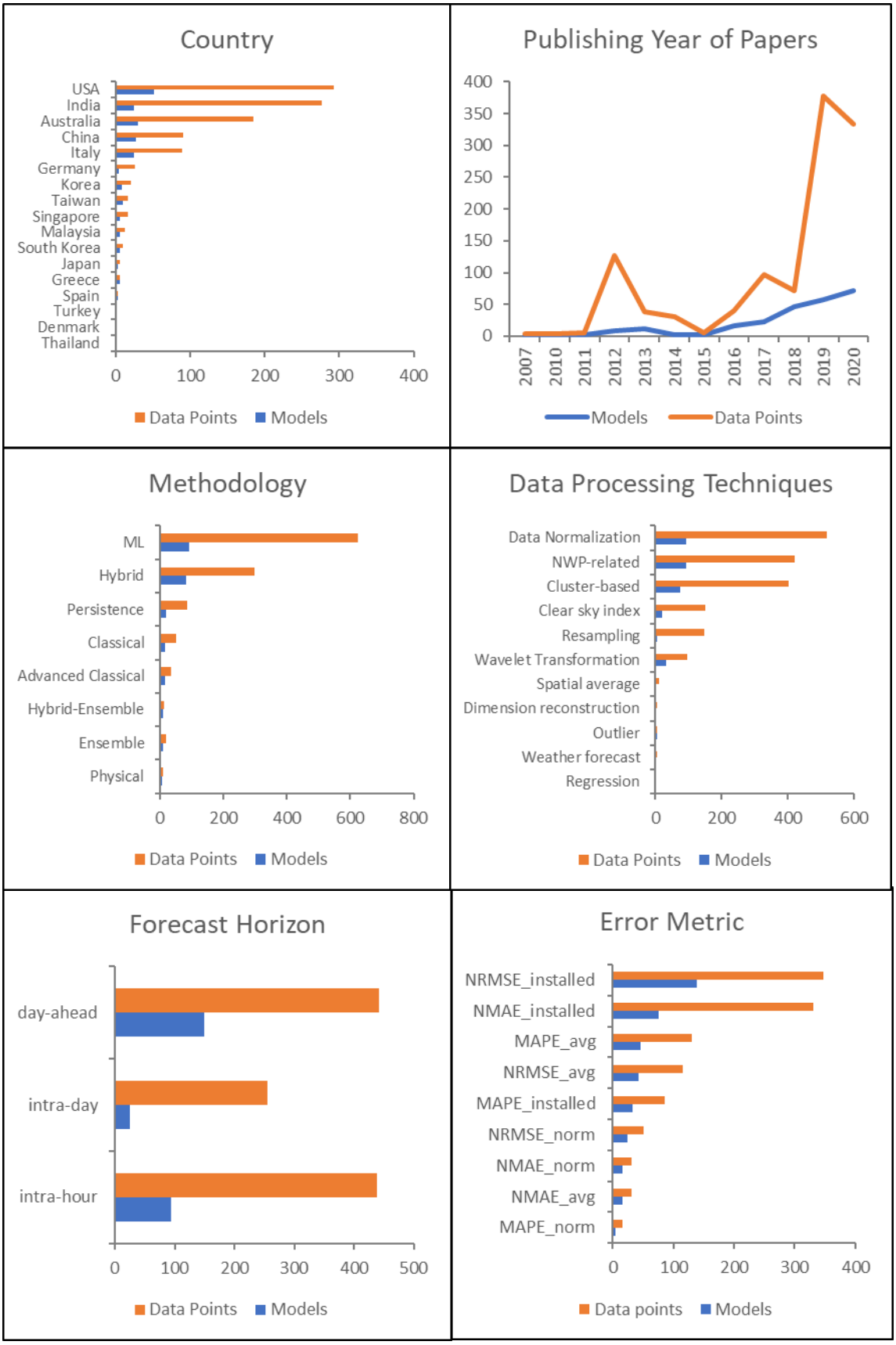

Figure 2: Data description

Figure 2 shows the distribution of data over 6 key factors. For each factor, the number of data points (the average error reported by a model in a paper) and models are counted.

As for the location of the data, our data base covers 74 regions across 17 countries and 4 continents. As can be seen from the panel for Country from Figure 2, the top 5 countries that cover most of the data points are the USA, India, Australia, China, and Italy. Overall, besides the USA and Australia, there are 9 Asian countries and 6 European countries.

Moving to the Publishing Year of Papers, we see an exponential increment in the number of models and data for PV output forecasts throughout the time, corresponding to the dominant role of PV in the global power supply. If we then look at the data distribution based on Methodology (see Appendix B for details on how this paper classify models), we observe that the state-of-the-art methodologies such as ML and hybrid methods dominate with 71% of total number of models and 81% of all data points. Other complex models including ensemble and hybrid-ensemble are only recently proposed and make up a small proportion. However, we show later that these models perform particularly well, and the future of PV output forecasts will be driven by such complex models.

The panel of Data Processing Techniques reveals which techniques are applied more frequently. As can be seen, the top candidates are data normalization, the inclusion of NWP variables, and cluster-based algorithms with 23%-30% of all observations for each technique, followed by clear sky index (9%), wavelet transformation (8%), and resampling (5%). For the other techniques, each accounts for less than 1% of all observations.

The lowest two panels describe the data distribution by forecast horizon and error metric. The left panel shows that the number of data points is higher for day-ahead and intra-hour, which reflects the fact that more effort has been driven to these two horizons, especially day-ahead forecasts with the highest number of models. And from the right panel, error metric divides the data base into 9 subsets corresponding to 3 error metrics and 3 error normalization methods. As can be seen, 89% of all data points concentrate on the top 5 error metrics and the subsets using normalized data to calculate errors contain particularly low number of observations. Details on the formulas of the error metrics are presented in Appendix C.

By now we have given an overview of the process of the statistical analysis on PV output forecasts. The next section discusses the results.

4 Results – What drives the accuracy of PV output forecasts?

Following we discuss each variable's effect on the PV forecast errors, starting with the length of the test set, followed by the forecast horizon, the time publishing the model, the complexity of the model (and the role of different data processing techniques), and the type of model (or methodology). For each variable, we begin with its coefficient in the OLS regressions and further explore its effect using data visualization methods.

The main regression (Equation (1))'s results are summarised in Table 3. As can be seen from this table, the dependent variable is the error value, and the explanatory variables include the test set length (days), the dummies of forecast horizons, the publishing year of the paper, the complexity of the model, and the dummies of methodologies. Column (1) and

(2) present the regression on the whole data base without any restriction and then restricting only observations of test sets at least one year. Column (3) and (4) compare the regression results between the classical and state-of-the-art methods.

Table 3: Factors influencing the accuracy of PV output forecasts

	Dependent variable: error values			
	Whole data base	Test sets >= 1 year (long test sets)		
	All methodologies (1)	All methodologies (2)	Classical models (3)	State-of-the-art models (4)
Test set length (days)	0.008***	0.010***	0.026***	0.007***
	(0.001)	(0.002)	(0.005)	(0.002)
Intra-day(1)	1.430*	3.445***		3.116***
	(0.747)	(0.834)		(0.825)
Day-ahead(1)	0.421	6.120***	7.720***	5.912***
	(0.652)	(0.865)	(2.371)	(0.922)
Publishing Year	-0.832***	-0.788***	-0.641	-0.976***
	(0.111)	(0.162)	(0.440)	(0.177)
Complexity	-0.340	-1.249***	0.371	-1.321***
	(0.234)	(0.379)	(1.106)	(0.407)
Classical(2)	-1.633	-0.906	0.784	
	(2.045)	(2.264)	(2.593)	
Ensemble(2)	1.840	0.923		
	(2.608)	(2.132)		
Hybrid(2)	-3.410**	-3.934**		-4.899***
	(1.629)	(1.667)		(1.547)
Hybrid-Ensemble(2)	-1.022	-0.568		-0.969
	(3.029)	(2.878)		(2.715)
ML(2)	-0.451	1.579		0.308
	(1.571)	(1.729)		(1.617)
Physical(2)	6.696**	-2.156	-0.385	
	(3.269)	(2.866)	(3.211)	
Constant	1,686.488***	1,594.978***	1,284.590	1,976.310***
	(224.071)	(327.278)	(887.429)	(356.840)
Observations	1,136	389	54	335
R^2	0.162	0.373	0.510	0.370
Adjusted R^2	0.153	0.353	0.436	0.355
Residual Std. Error	8.991 (df = 1123)	5.792 (df = 376)	6.069 (df = 46)	5.652 (df = 326)
F Statistic	18.057*** (df = 12; 1123)	18.631*** (df = 12; 376)	6.850*** (df = 7; 46)	23.957*** (df = 8; 326)

Note: (1) Dummies of forecast horizon, baseline: intra-hour horizon
(2) Dummies of methodology, baselines: column (1-3): Advanced classical models, column (4): Ensemble models
*p<0.1; **p<0.05; ***p<0.01

Now let us discuss each variable's effects on the PV forecast errors in detail.

4.1 Test set length

First let us discuss the effects of the test set length on the accuracy of PV output forecast models looking at its coefficient from the regression in Table 3. As can be seen, the coefficients are highly statistically significant and positive (0.007-0.026) for all cases, indicating a positive correlation of error values with the length of the test sets. More

interestingly, this correlation is much stronger for the classical methods than the state-of-the-art, with each additional day in the test set leading to an error increase of 0.026 percentage point (pp) for the former and only 0.007 pp for the latter.

Regarding the test set length variable, we also examine the importance of the long test sets and the "cherry picking" hypothesis as suggested by many scholars discussed in Section 2.1. On Table 3, moving from column (1) (regression on the whole data base) to column (2) (regression on only the observations of test sets at least one year), we see that the coefficients of most variables have larger magnitudes and become more significant, with the explanation power of the variables (adjusted R2) increasing from 15% to 35%. This indicates that requiring the test sets to be at least one year allows the data base to generate more meaningful results, which supports the argument of many scholars that longer test sets can test the robust performance of models.

The "cherry picking" hypothesis is verified by comparing the errors reported on one single day and the other test sets. As can be seen from Figure 3, the one single day test sets have a remarkably lower errors compared to the other test sets. Pulling all the error metrics, one-day test sets have the average error value of 2.7%, which is around a quarter of that of all the other test sets (~10%).

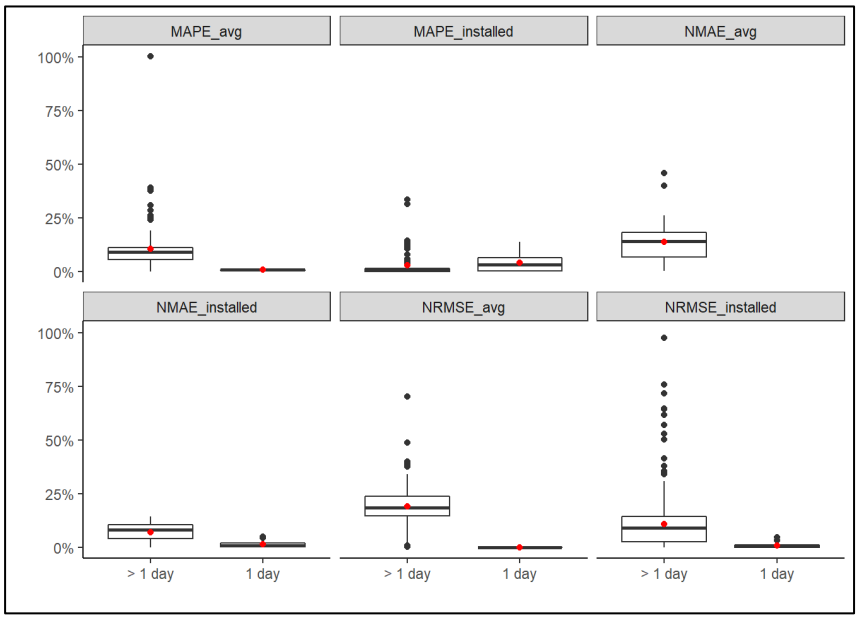

Figure 3: Error values with length of test sets
Note: The "> 1 day" test sets have the average MAE and RMSE normalized by installed capacity or peak power (NMAE_installed and NRMSE_installed) 5-10 times higher than the "1 day" test sets. The gap is up to 641 times when considering the RMSE normalized by average or measured values (NRMSE_avg), with the "1 day" test sets having the average error of only 0.03%.

The significant gap of errors between the one-day test sets and the others remains robust when we further remove the risks of bias caused by the methodology and forecast horizons. As can be seen from Figure 4, the one-day test sets achieve consistently lower errors than the other test sets for all groups of models and forecast horizons. This implies the possibility of "cherry picking" in reporting errors and emphasizes the necessity of having a long (and standardised if possible) test set in assessing models' performance.

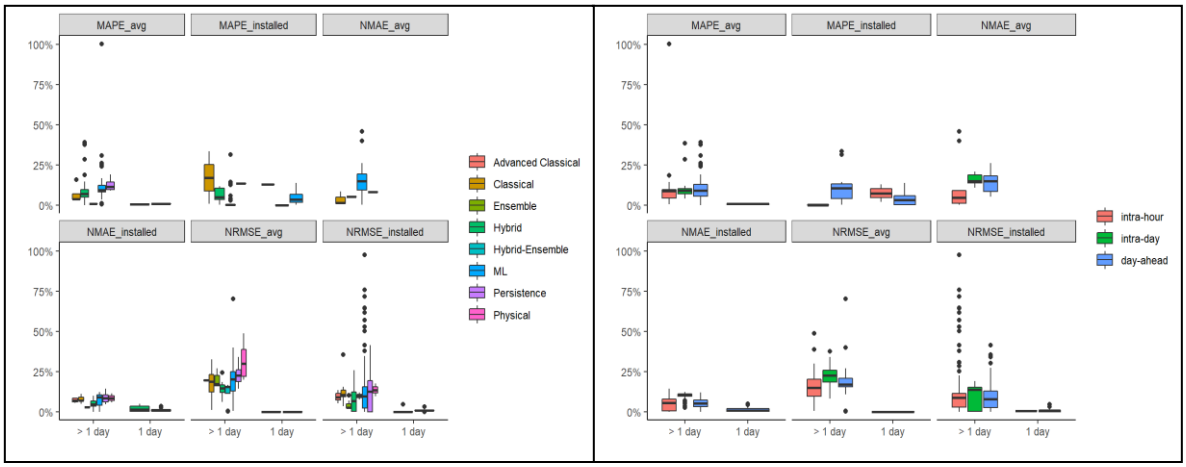

Figure 4: Error values with different lengths of test sets, grouped by methodologies (left panel) and forecast horizons (right panel)

4.2 Forecast horizon

The next variable is the forecast horizons. The coefficients of the dummies of the forecast horizons in Table 3 show that changing from intra-hour forecasts (baseline) to longer horizons such as intra-day and day-ahead increases the average errors remarkably. In the data restricting the bias caused by the test set length (column (2)), the intra-day and day-ahead forecast errors are higher than the intra-hour by 3.45 pp and 6.12 pp respectively. Classical models seem more sensitive to the change in the forecast horizon than the state-of-the-art methods, with the intra-hour–day-ahead error gap being 7.72 pp for the former compared with 5.91 pp for the later, implying the superiority of state-of-the-art methodologies in long horizon forecasts.

Let us now further explore the effects of forecast horizons harmonising the other context difference.

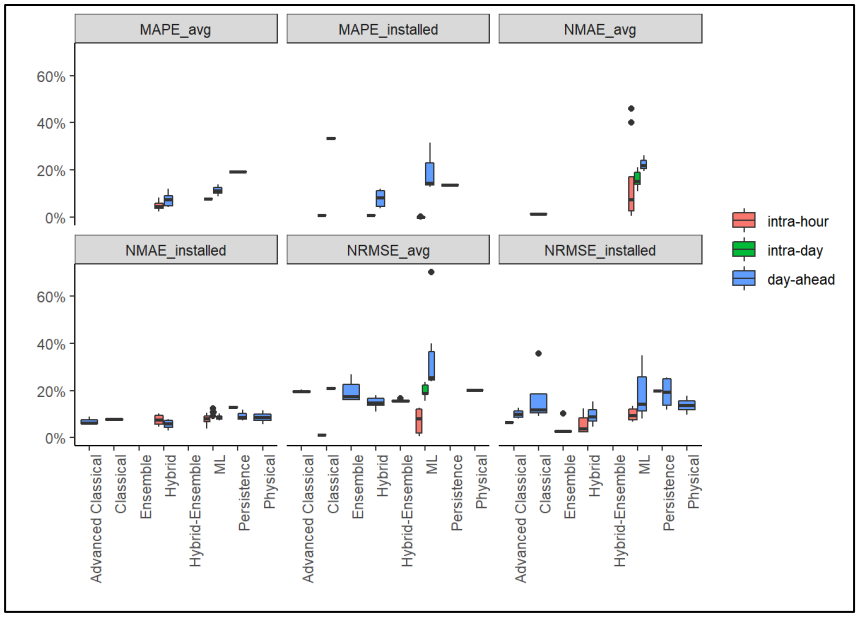

Figure 5: Error values with forecast horizons

Figure 5 compares the forecast errors between the forecast horizons in each group of methodologies and error metrics, using the data of long test sets (at least one year). The figure clearly confirms the positive correlation between the error and the length of the forecast horizons. Looking at the ML methods, for example, we see a remarkable increase in the values of NMAE_avg and NRMSE_avg when we move from intra-hour to intra-day and then to day-ahead forecasts.

This part confirms the suggestions made by many scholars that the longer horizons are, the more difficult it is to have good accuracy of forecasts. This fact implies that more effort will be driven towards improving long horizon forecasts such as the day-ahead. Forecast horizon is indeed a very important factor deciding the relative performance of models and should be taken into account in all analysis of PV output forecasts.

4.3 Time of publishing the papers

As for the publishing year of papers, it shows a significantly negative correlation with the forecast errors. The regressions in Table 3 show that models published one year later have the average errors that are 0.64-0.98 pp lower (column (3) and (4)). The coefficient is highly statistically significant for state-of-the-art models while showing no significance for classical models, indicating a more consistent improvement in forecast quality of the state-of-the-art models compared with the classical ones, though the overall effect observed for all methodologies is negative.

The overall improvement in the forecast accuracy is also reflected in Figure 6, which shows a huge success in lowering the errors of the PV output forecasts from 2007 to 2020. On average, there was a decrease of 2 pp annually, bringing the average error value from 35% in 2007 to less than 8% in 2020.

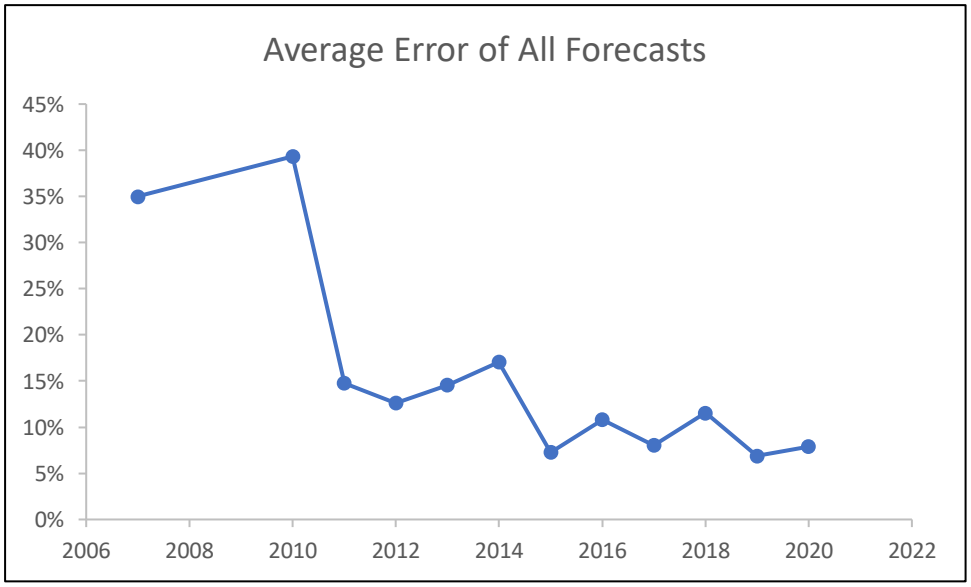

Figure 6: PV output forecast progress

We take a further step to examine the progress of PV output forecast accuracy taking into account the other risks of bias. As observed from Figure 7, all error metrics, methodologies

and forecast horizons show a downward trend in the error values, allowing a concrete conclusion on the forecast errors decreasing with time.

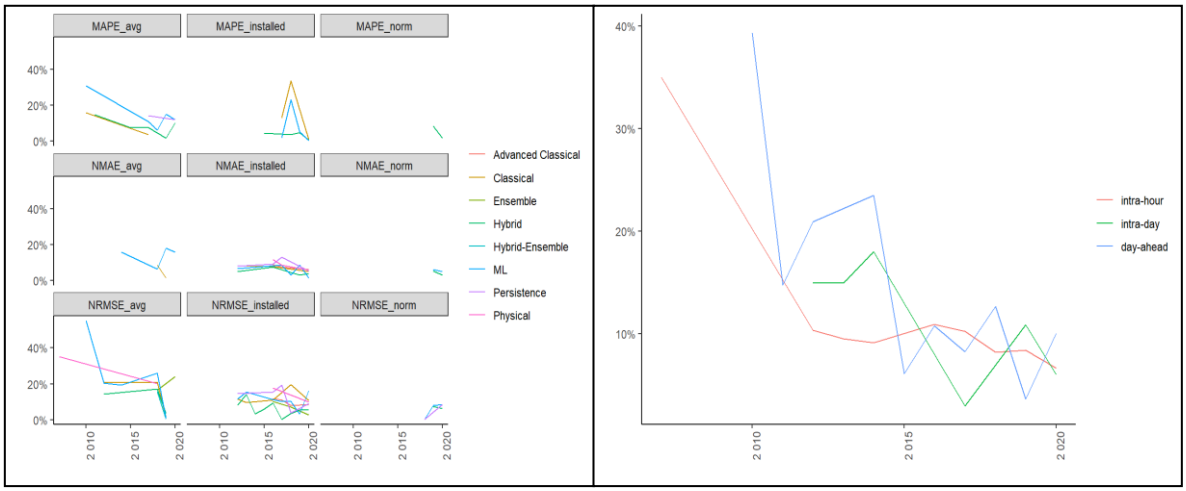

Figure 7: PV output progress for different error metrics and methodologies (left panel) and forecast horizons (right panel)

4.4 Complexity of models and the use of data processing techniques

The statistical analysis of the data base proves that the complexity of the models, which counts the number of data processing techniques used by the models, is negatively correlated with the forecast errors, especially significant for the state-of-the-art methodologies.

The regression presented in Table 3 shows that each one additional technique reduces the average errors by 1.25 pp for the pool of all models (column (2)) and by 1.32 pp for the state-of-the-art methods (column (4)). However, the complexity variable has no statistically significant effect on the group of classical methods. Figure 8 visualizes the effects of different level of model's complexity on the average error in each error metric and methodology group.

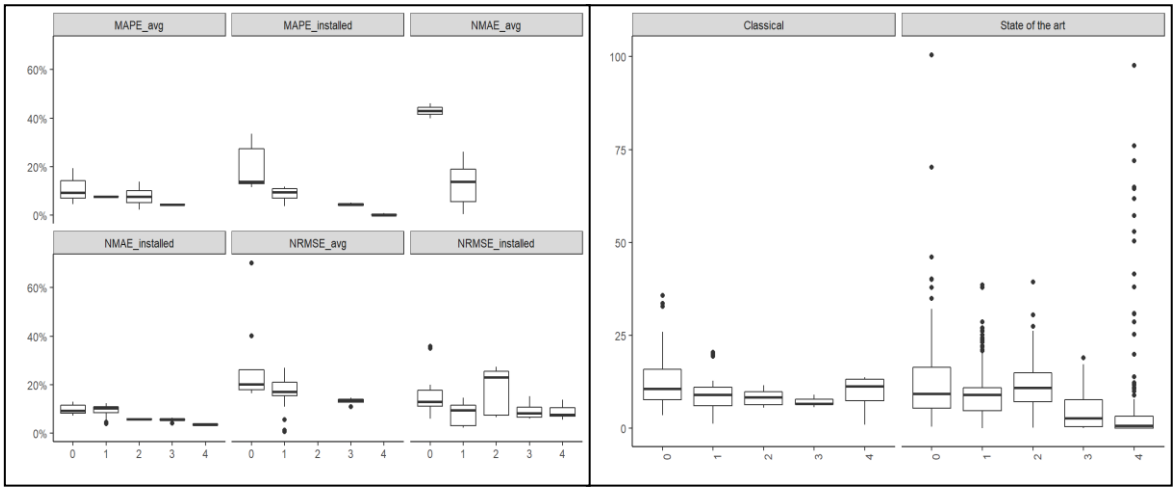

Figure 8: Error values with complexity of models for each error metric (left panel) and methodology (right panel)

As can be seen, models using 0-2 techniques have the average errors being 44.7% higher compared with those using 3-4 techniques. This gap can be up to 99% for state-of-the-art methods, while the pattern is not so clear for the classical models, indicating that data processing techniques have much stronger influence on state-of-the-art models (than the classical).

In addition, we also examine the comparative efficiency of individual data processing techniques, using the modified version of the main regression (Equation (2)). Table 4 presents the regression results, with each column reporting the effects of each data processing technique on the forecast errors, controlling for the variables of test set length, forecast horizon, publishing year of the model, types of models, and the effects of other data processing techniques. As can be seen, the technique of data normalization is the most effective, reducing the average error of the model by 3.16 pp, followed by resampling technique (-2.88 pp) and the inclusion of NWP model's output (-2.48 pp). These are also among the most used techniques as discussed in Section 3.2 (Figure 2). More interestingly, although cluster-based and WT are also frequently used in data processing, these techniques do not show significant influence on the forecast accuracy.

Our analysis on data processing techniques has not only confirmed the crucial importance of these techniques to the performance of PV output forecasts. More importanly, we show where these techniques can function the best and which ones are the most effective. These findings are essential for the further improvement of PV output forecast accuracy.

Table 4: Effects of data processing techniques on error values

	Dependent variable: error value										
	Cluster-based (1)	NWP-related (2)	Normalization (3)	WT (4)	Outlier (5)	CSI (6)	Spatial average (7)	Resampling (8)	Weather forecast (9)	Regression (10)	Dimension Reconstruction (11)
Processing technique	0.939 (1.352)	**-2.478*** (1.297)	**-3.162***** (0.755)	-0.977 (1.227)	-5.482 (4.294)	3.259*** (1.144)	0.632 (4.252)	**-2.877**** (1.205)	-1.647 (4.576)	-5.667 (9.007)	-0.944 (3.822)
Test set length (days)	0.008*** (0.001)	0.008*** (0.001)	0.008*** (0.001)	0.008*** (0.001)	0.008*** (0.001)	0.008*** (0.001)	0.008*** (0.001)	0.008*** (0.001)	0.008*** (0.001)	0.008*** (0.001)	0.008*** (0.001)
Intra-day[1]	2.228*** (0.825)	2.228*** (0.825)	2.228*** (0.825)	2.228*** (0.825)	2.228*** (0.825)	2.228*** (0.825)	2.228*** (0.825)	2.228*** (0.825)	2.228*** (0.825)	2.228*** (0.825)	2.228*** (0.825)
Day-ahead[1]	1.253 (0.787)	1.253 (0.787)	1.253 (0.787)	1.253 (0.787)	1.253 (0.787)	1.253 (0.787)	1.253 (0.787)	1.253 (0.787)	1.253 (0.787)	1.253 (0.787)	1.253 (0.787)
Publishing Year	-0.698*** (0.134)	-0.698*** (0.134)	-0.698*** (0.134)	-0.698*** (0.134)	-0.698*** (0.134)	-0.698*** (0.134)	-0.698*** (0.134)	-0.698*** (0.134)	-0.698*** (0.134)	-0.698*** (0.134)	-0.698*** (0.134)
Classical[2]	-2.148 (2.046)	-2.148 (2.046)	-2.148 (2.046)	-2.148 (2.046)	-2.148 (2.046)	-2.148 (2.046)	-2.148 (2.046)	-2.148 (2.046)	-2.148 (2.046)	-2.148 (2.046)	-2.148 (2.046)
Ensemble[2]	0.506 (3.466)	0.506 (3.466)	0.506 (3.466)	0.506 (3.466)	0.506 (3.466)	0.506 (3.466)	0.506 (3.466)	0.506 (3.466)	0.506 (3.466)	0.506 (3.466)	0.506 (3.466)
Hybrid[2]	-3.534** (1.679)	-3.534** (1.679)	-3.534** (1.679)	-3.534** (1.679)	-3.534** (1.679)	-3.534** (1.679)	-3.534** (1.679)	-3.534** (1.679)	-3.534** (1.679)	-3.534** (1.679)	-3.534** (1.679)
Hybrid-Ensemble[2]	-0.923 (3.298)	-0.923 (3.298)	-0.923 (3.298)	-0.923 (3.298)	-0.923 (3.298)	-0.923 (3.298)	-0.923 (3.298)	-0.923 (3.298)	-0.923 (3.298)	-0.923 (3.298)	-0.923 (3.298)
ML[2]	-0.390 (1.650)	-0.390 (1.650)	-0.390 (1.650)	-0.390 (1.650)	-0.390 (1.650)	-0.390 (1.650)	-0.390 (1.650)	-0.390 (1.650)	-0.390 (1.650)	-0.390 (1.650)	-0.390 (1.650)
Physical[2]	5.894* (3.297)	5.894* (3.297)	5.894* (3.297)	5.894* (3.297)	5.894* (3.297)	5.894* (3.297)	5.894* (3.297)	5.894* (3.297)	5.894* (3.297)	5.894* (3.297)	5.894* (3.297)
Constant	1,418.008*** (270.382)	1,418.008*** (270.382)	1,418.008*** (270.382)	1,418.008*** (270.382)	1,418.008*** (270.382)	1,418.008*** (270.382)	1,418.008*** (270.382)	1,418.008*** (270.382)	1,418.008*** (270.382)	1,418.008*** (270.382)	1,418.008*** (270.382)

Observations	1,136	1,136	1,136	1,136	1,136	1,136	1,136	1,136	1,136	1,136	1,136
R^2	0.175	0.175	0.175	0.175	0.175	0.175	0.175	0.175	0.175	0.175	0.175
Adjusted R^2	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159
Residual Std. Error	8.957 (df = 1113)	8.957 (df = 1113)	8.957 (df = 1113)	8.957 (df = 1113)	8.957 (df = 1113)	8.957 (df = 1113)	8.957 (df = 1113)	8.957 (df = 1113)	8.957 (df = 1113)	8.957 (df = 1113)	8.957 (df = 1113)
F Statistic	10.766*** (df = 22; 1113)	10.766*** (df = 22; 1113)	10.766*** (df = 22; 1113)	10.766*** (df = 22; 1113)	10.766*** (df = 22; 1113)	10.766*** (df = 22; 1113)	10.766*** (df = 22; 1113)	10.766*** (df = 22; 1113)	10.766*** (df = 22; 1113)	10.766*** (df = 22; 1113)	10.766*** (df = 22; 1113)

Note: (1) Dummies of forecast horizon, baseline: intra-hour horizon
(2) Dummies of methodology, baseline: Advanced classical models
*p<0.1; **p<0.05; ***p<0.01

4.5 Type of models - Which methodology is superior?

So far, we have discussed the effects of four important variables on the PV output forecast errors. In this part, we study the role of the most important factor – the type of models or the methodology.

First let us get back to the regression in Table 3. The coefficients of methodology dummies show that hybrid methods consistently achieve significantly lower errors than the other models, reducing the average errors by from 3.41-3.93 pp compared to the advanced classical methods (column (1)) and by 4.90 pp compared to the ensemble models (column (4)). The other methodologies do not show statistically significant influence on error values for almost all the cases. Although no clear rank is made for all types of models, the regression results indicate the dominant position of the hybrid models in PV output forecasts.

We further explore the inter-methodology comparison using boxplot method. Figure 9 compares models' errors using different error metrics, with the left panel presenting the whole data base and the right one restricting the bias from test set length by keeping only at least one year test sets. Interestingly, the findings from the left panel reflect the opinions and summary of many historical review papers on PV output forecast models' performance, while the right panel examines these opinions more critically and unbiasedly.

On the left panel, the state-of-the-art methodologies including the hybrid, ensemble, hybrid-ensemble, and ML methods are the best candidates, though ML's performance is varied with a large number of outliers and a large gap between the best and the worst forecasts. Particularly, complex models, i.e., the hybrid, ensemble, and hybrid-ensemble models, outperform the individual ones substantially. Take day-ahead forecasts as an example, hybrid-ensemble methods report NRMSE_avg 32%-40% lower, and ensemble methods report NRMSE_installed 57%-72% lower than the individual models. These results support the arguments for the superiority of the state-of-the-art models.

However, the error gap above between the complex and individual models is much lower when we look at the right panel of Figure 9. Focusing on the day ahead forecasts (as the other two horizons are left with too few observations when keeping only long test sets), we see that even though the hybrid, ensemble and hybrid-ensemble methods still achieve the lowest average error values for all error metrics, these models achieve the errors that are 9%-24% lower than that of the individual methods instead of up to 72% as observed on the left side. This indicates that comparing models' performance without considering the bias effects of other factors (e.g., the test set length) can lead to misleading conclusions, and particularly in this case, can overestimate the achievement of complex models.

The bias is also observed for the assessment of ML models. While ML technique has almost equal average errors to the other state-of-the-art methods on the left panel of Figure 9, its comparative performance is significantly lower on the right panel when restricting the bias of test set length. For example, ML's average NRMSE_avg for day-ahead forecasts on the left panel is 17.5%, which is close to the performance of hybrid-ensemble models (12%). However, the same error metric increases to 35% for ML models on the right panel, which more than doubles that of any complex methods and is 67%-77% higher than simple and advanced classical methods. In many cases, ML can show very bad performance, even when compared with the persistence models.

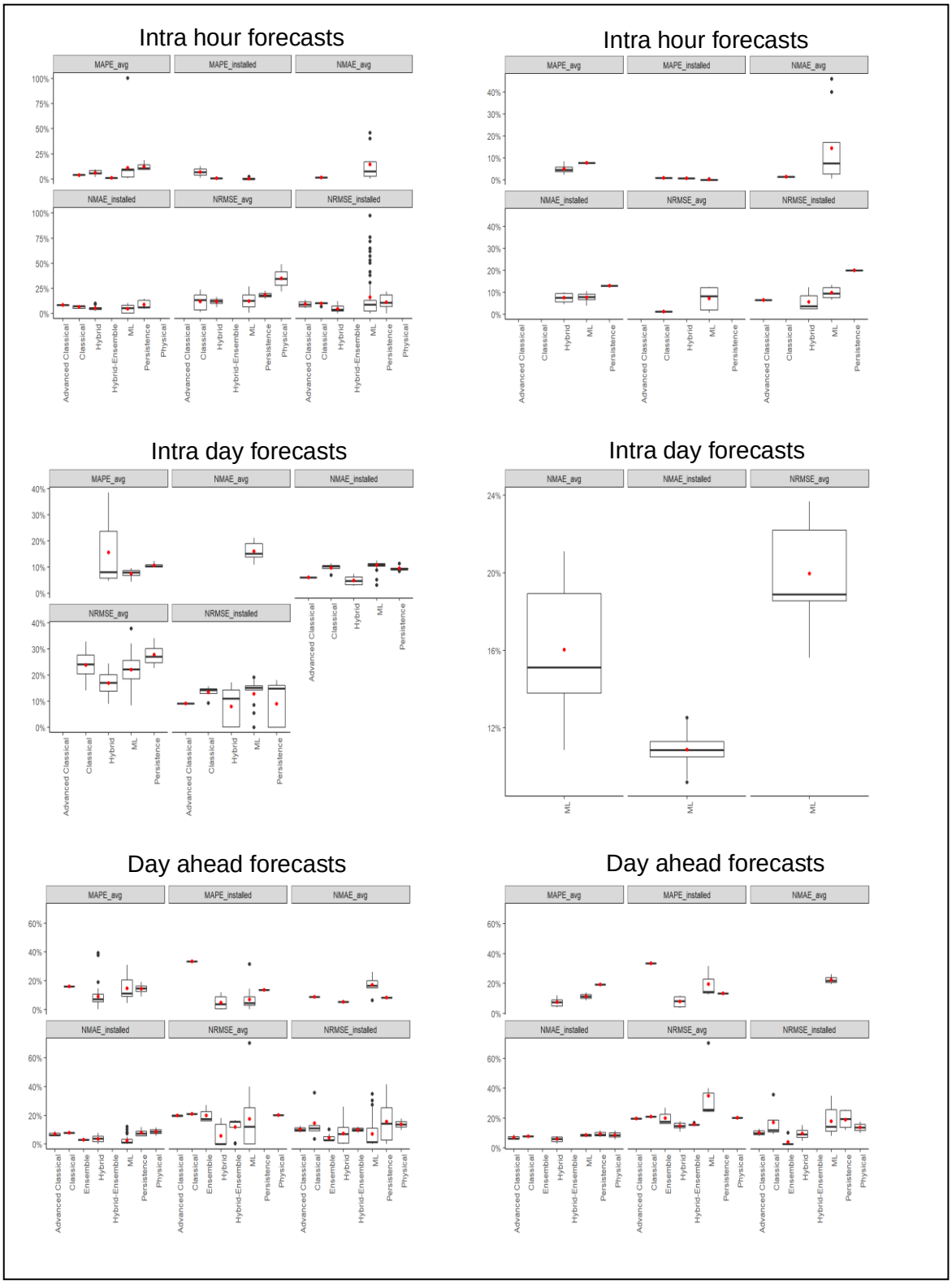

Figure 9: Methodologies' comparative performance using the data base (left panels) and restricted data set (right panels)

Therefore, considering the computational burden of the state-of-the-art methods, classical methods can be a better choice to balance the accuracy and the costs for the forecasts in the short and medium term. In the long term, however, we show below that it is worth investing in state-of-the-art methodologies to further improve the PV output forecasts.

Indeed, Figure 10 compares the progress of these two groups of methodologies, using the average error value (the left graph) and the minimum error value (the right graph). From the left graph, we see that although the classical methods beat the state-of-the-art at some points, the overall progress made in the state-of-the-art group is much larger, with an annual decrease of 3.94 pp during the last 10 years, compared with 0.94 pp for classical methods. On the right graph, we show that there is a large gap in the minimum error value that can be achieved by state-of-the-art methods compared to the classical, indicating much potential of state-of-the-art methods in reducing the forecast errors.

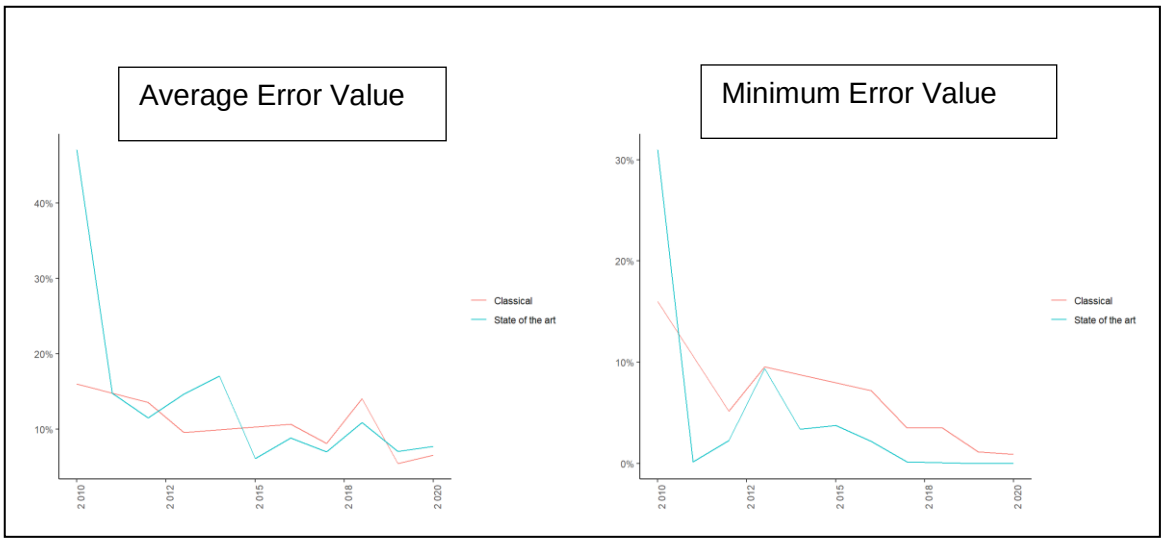

Figure 10: A comparison of classical and state-of-the-art models' progress

Our analysis in section 4.4 of the stronger influence of data processing techniques on state-of-the-art models than on the classical can partially explain for the higher potential of the state-of-the-art methods. Though being less dependent on extra techniques makes the classical models' performance more stable, it also means less likelihood to have leaping improvements. Considering increasing effort driven towards improving the data processing and optimization techniques in the past years, the state-of-the-art methods have high potential to further enhance the forecast accuracy in the long run.

So far, the analysis of the data allows us to safely confirm the superiority of the complex models such as the hybrid, ensemble, and hybrid-ensemble methods, though much less impressive when removing some important sources of bias. ML models, though assessed as performing well by many scholars, do not show a robustly good performance. However, the state-of-the-art methodologies have more potential to further improve forecast quality in the long run compared with the classical methods.

Noteworthily, methodology comparison requires a data base that is large enough to allow harmonising different features and still leave enough observations for comparison. This requires huge effort to collect, process and analyse the data. In the meanwhile, new PV forecast models are proposed every day, which requires a continuous update of the data base and thus costs more resources. A benchmark is therefore the most efficient and easiest way to systematize knowledge and compare models. This leads us to the final section.

5 A benchmark for PV output forecast assessment

An established benchmark for PV output forecasts has numerous advantages. First, a benchmark ensures that all models are tested in an identical transparent context and use the same error report methods, which allows direct comparison of error values among models. Second, a benchmark is also an open space that benefits both scholars and investors. As for scholars, a benchmark puts them on a transparent playground and diminishes all context preferences and the risks of bias, which motivates more competition and thus faster progress. Furthermore, the scholars can easily and quickly track their ranks among the community, which is pivotally important for further improvements in PV output forecasts. For investors, having the plant's data as among the standardised data sets used for the benchmark allows them to use the resources from scholars all over the world, who can contribute to enhancing the forecast accuracy for the investors' PV plant "for free". More importantly, a benchmark is a dynamic and open space, where models' performance and rankings are updated continuously as new models are tested without demanding any additional effort to collect and update data. The participation of a variety of methodologies and data sets also facilitates the transfer learning in PV output forecast domain and contributes enormously to the accuracy improvement.

Therefore, we suggest the following steps to establish a benchmark:

(i) Have a standardised suit of evaluation metrics with formal requirements and instructions for the error reporting process.

As discussed above, there are a large number of options to report the forecast errors, which means fewer data points in each error metric group and causes difficulty in comparing models. Therefore, the evaluation metrics must be standardised.

Among the error metrics, we recommend MAE and RMSE to assess the forecast quality for both long and short terms. MAE, with its focus on mean error values, is less sensitive to variability of the data set and is more suitable to long-term forecasts for management and planning purposes. As for RMSE, the squared values make it more sensitive to outliers and spikes in data (e.g., severe solar ramps), therefore satisfying the key requirement in short term PV forecasts – capturing the model's forecast accuracy in extreme events (Blaga *et al.*, 2019).

Besides, as many scholars also attempted to suggest new approaches of evaluating the models' performance, arguing that one single metric cannot represent the whole model (e.g., Marquez and Coimbra (2013)), in addition to the above suggested error metrics, the benchmark could periodically include more new metrics to the standardised suit to promote the fair assessment.

More importantly, we also mention above that scholar can have different calculation approaches for a same error metric, or simply report normalized errors without defining their calculation mechanism or the reference quantity for error normalization. To solve this problem, formal instructions and requirements should be made on the model testing progress to ensure the transparency in model assessment.

(ii) Have a bank of standardised data sets for training and testing models.

The next step would be to have standardised data sets to eliminate all context difference in models training and testing. Any investors or scholars who would like to contribute to the bank of data sets can send their data sets to the benchmark coordinator to be examined and standardised. In this way, the bank of data sets will always be kept updated.

(iii) Have an open space for the benchmark.

Finally, a benchmark should be established as an open space, preferably by leaders of both scholar and industry community, so that it can be accepted, widely used, and contributed to by many scholars, which is the prerequisite for the success of the benchmark. The benchmark can be initiated as competitions in the beginning to attract scholars to participate in. In the long run, quarterly or annual rankings can be made for the models, which not only informs all stakeholders about the progress in PV output forecast, but also attracts more participation from scholars and industries, leading to the further development of the benchmark – the systematic data base of PV output forecast assessment.

6 Conclusion

Accurate photovoltaic (PV) forecasts are increasingly important to the integration of PV into grid, attracting a consistently high interest from grid operators, investors, politicians, and forecasters from both industry and academia. This leads to such a vast number of literatures focusing on enhancing PV forecast accuracy that it requires a scientific knowledge systemization.

While there are already some survey papers summarizing findings from the literature, our work is the first statistical analysis on PV output forecasts to concretely answer the question “What drives the accuracy of PV output forecasts?”. To do that, we examine all the literature on PV output forecasts that we could find, assess their quality, extract the data from the papers and build a data base of models’ forecast errors including 1,136 observations with 21 key features, covering a variety of models, regions, training and testing data sets etc, which is large enough to control for the risks of bias from various factors and produce robust, statistically significant results.

Using OLS regression and data visualization methods to analyse the data base, we come up with the following conclusions:

- Out-of-sample test set length positively correlates with the forecast errors. An additional day in the test set increases the error by 0.007-0.026 pp. The effect is larger for the classical models than for the state-of-the-art models, indicating a more robust performance of the latter.
- Long test sets (at least one year) generate more meaningful conclusions on PV output forecast assessment. Restricting the bias from the difference in test set lengths can double the explanation power of the regression from 15% to 35%.
- The possibility of “cherry picking” in reporting errors exists. One-day test sets have the average error value of 2.7%, which is around a quarter of that of all the other test sets (~10%).
- The longer the forecast horizons are, the more difficult to have high forecast accuracy. On average, the intra-day and day-ahead forecast errors are higher than

the intra-hour by 3.45 pp and 6.12 pp respectively. The classical models are more sensitive to the change in forecast horizons than the state-of-the-art, implying the high potential of the state-of-the-art methods in improving the long horizon forecasts.

- PV output forecasts have a steady improvement. Models published one year later have the average errors that are 0.64-0.98 pp lower. The progress is more significant for the state-of-the-art than for the classical methods.
- Data processing techniques contributes to enhancing the forecast accuracy. Each one additional technique reduces the average errors by 1.25-1.32 pp. The effect is stronger for state-of-the-art methods, signalling the further improvement that can be made in the long run by this group of methodologies.
- Among the data processing techniques observed in the data base, the technique of data normalization is the most effective, reducing the average error of the model by 3.16 pp, followed by resampling technique (-2.88 pp) and the inclusion of NWP model's output (-2.48 pp).
- Hybrid, ensemble, and hybrid-ensemble models achieve the lowest forecast errors. Hybrid models are consistently superior to the others and outperform the classical methods by 3.41-3.93 pp. Hybrid-ensemble methods also achieve the NRMSE_avg that is 32%-40% lower, and ensemble methods report NRMSE_installed that is 57%-72% lower than the individual models.
- In the meanwhile, ML performs much worse when removing the key risks of bias in inter-model comparison. For example, analysing the data base of all test set lengths, ML's average NRMSE_avg for day-ahead forecasts is 17.5%, which is close to the performance of hybrid-ensemble models (12%). However, when we include only the test sets of at least one year length, the same error metric increases to 35% for ML models– compared with hybrid methods (15-17%) and classical methods (19-20%).
- The superiority of the state-of-the-art methods can be overestimated if we do not consider the risks of bias caused by context difference. The complexity-accuracy trade-off therefore favours the classical models in the short and medium run. However, the complex models show much higher potential to enhance forecasts' quality in the long run thanks to the development of new data processing techniques. The future of PV output forecasts is consequently driven by the state-of-the-art models.

These findings, as important materials for scholars to inherit systematic knowledge from historical literature, are crucial to the future development of PV output forecasts. Through the analysis process, we also realize how costly it is to conduct such a statistical analysis on a huge number of papers, which can be saved through a benchmark for assessing PV output forecasts. This paper takes the very first step towards establishing this benchmark.

Literatur

Acharya, S.K., Wi, Y.-M. and Lee, J. (2020), “Day-Ahead Forecasting for Small-Scale Photovoltaic Power Based on Similar Day Detection with Selective Weather Variables”, Electronics, Vol. 9 No. 7, p. 1117.

Ahmed, R., Sreeram, V., Mishra, Y. and Arif, M.D. (2020), “A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization”, Renewable and Sustainable Energy Reviews, Vol. 124, p. 109792.

Akhter, M.N., Mekhilef, S., Mokhlis, H. and Mohamed Shah, N. (2019), “Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques”, IET Renewable Power Generation, Vol. 13 No. 7, pp. 1009–1023.

Almonacid, F., Pérez-Higueras, P.J., Fernández, E.F. and Hontoria, L. (2014), “A methodology based on dynamic artificial neural network for short-term forecasting of the power output of a PV generator”, Energy Conversion and Management, Vol. 85, pp. 389–398.

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de-Pison, F.J. and Antonanzas-Torres, F. (2016), “Review of photovoltaic power forecasting”, Solar Energy, Vol. 136, pp. 78–111.

Asrari, A., Wu, T.X. and Ramos, B. (2017), “A Hybrid Algorithm for Short-Term Solar Power Prediction—Sunshine State Case Study”, IEEE Transactions on Sustainable Energy, Vol. 8 No. 2, pp. 582–591.

Baharin, K.A., Abdul Rahman, H., Hassan, M.Y. and Gan, C.K. (2016), “Short-term forecasting of solar photovoltaic output power for tropical climate using ground-based measurement data”, Journal of Renewable and Sustainable Energy, Vol. 8 No. 5, p. 53701.

Barbieri, F., Rajakaruna, S. and Ghosh, A. (2017), “Very short-term photovoltaic power forecasting with cloud modeling: A review”, Renewable and Sustainable Energy Reviews, Vol. 75, pp. 242–263.

Blaga, R., Sabadus, A., Stefu, N., Dughir, C., Paulescu, M. and Badescu, V. (2019), “A current perspective on the accuracy of incoming solar energy forecasting”, Progress in energy and combustion science, Vol. 70, pp. 119–144.

Borenstein, M., Hedges, L.V., Higgins, J.P.T. and Rothstein, H.R. (2009), Introduction to Meta-Analysis, John Wiley & Sons, Ltd, Chichester, UK.

Bouzerdoum, M., Mellit, A. and Massi Pavan, A. (2013), “A hybrid model (SARIMA–SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant”, Solar Energy, Vol. 98, pp. 226–235.

Brereton, P., Kitchenham, B.A., Budgen, D., Turner, M. and Khalil, M. (2007), “Lessons from applying the systematic literature review process within the software engineering domain”, Journal of Systems and Software, Vol. 80 No. 4, pp. 571–583.

Chavez Velasco, J.A., Tawarmalani, M. and Agrawal, R. (2021), “Systematic Analysis Reveals Thermal Separations Are Not Necessarily Most Energy Intensive”, Joule, Vol. 5 No. 2, pp. 330–343.

Chen, B., Lin, P., Lin, Y., Lai, Y., Cheng, S., Chen, Z. and Wu, L. (2020), “Hour-ahead photovoltaic power forecast using a hybrid GRA-LSTM model based on multivariate meteorological factors and historical power datasets”, IOP Conference Series: Earth and Environmental Science, Vol. 431 No. 1, p. 12059.

Chen, C., Duan, S., Cai, T. and Liu, B. (2011), “Online 24-h solar power forecasting based on weather type classification using artificial neural network”, Solar Energy, Vol. 85 No. 11, pp. 2856–2870.

Chupong, C. and Plangklang, B. (2011), “Forecasting power output of PV grid connected system in Thailand without using solar radiation measurement”, Energy Procedia, Vol. 9, pp. 230–237.

Da Liu and Sun, K. (2019), “Random forest solar power forecast based on classification optimization”, Energy, Vol. 187, p. 115940.

Da Silva Fonseca, J.G., Oozeki, T., Takashima, T., Koshimizu, G., Uchida, Y. and Ogimoto, K. (2012), “Use of support vector regression and numerically predicted cloudiness to forecast power output of a photovoltaic power plant in Kitakyushu, Japan”, Progress in Photovoltaics: Research and Applications, Vol. 20 No. 7, pp. 874–882.

Dalton, G.J., Lockington, D.A. and Baldock, T.E. (2009), “Feasibility analysis of renewable energy supply options for a grid-connected large hotel”, Renewable Energy, Vol. 34 No. 4, pp. 955–964.

Dan A. Rosa De Jesus, Paras Mandal, Miguel Velez-Reyes, Shantanu Chakraborty and Tomonobu Senjyu (2019), “Data Fusion Based Hybrid Deep Neural Network Method for Solar PV Power Forecasting”, paper presented at 2019 North American Power Symposium (NAPS), available at: https://www.researchgate.net/publication/339328814_Data_Fusion_Based_Hybrid_Deep_Neural_Network_Method_for_Solar_PV_Power_Forecasting.

Das, U., Tey, K., Seyedmahmoudian, M., Idna Idris, M., Mekhilef, S., Horan, B. and Stojcevski, A. (2017), “SVR-Based Model to Forecast PV Power Generation under Different Weather Conditions”, Energies, Vol. 10 No. 7, p. 876.

Das, U.K., Tey, K.S., Seyedmahmoudian, M., Mekhilef, S., Idris, M.Y.I., van Deventer, W., Horan, B. and Stojcevski, A. (2018), “Forecasting of photovoltaic power generation and model optimization: A review”, Renewable and Sustainable Energy Reviews, Vol. 81, pp. 912–928.

Ding, M., Wang, L. and Bi, R. (2011), “An ANN-based Approach for Forecasting the Power Output of Photovoltaic System”, Procedia Environmental Sciences, Vol. 11, pp. 1308–1315.

Dokur, E. (2020), “Swarm Decomposition Technique Based Hybrid Model for Very Short-Term Solar PV Power Generation Forecast”, Elektronika ir Elektrotechnika, Vol. 26 No. 3, pp. 79–83.

E. Lorenz, D. Heinemann, Hashini Wickramarathne, H. Beyer and S. Bofinger (2007), “Forecast of ensemble power production by grid-connected pv systems”.

El hendouzi, A. and Bourouhou, A. (2020), “Solar Photovoltaic Power Forecasting”, Journal of Electrical and Computer Engineering, Vol. 2020, pp. 1–21.

Eseye, A.T., Zhang, J. and Zheng, D. (2018), “Short-term photovoltaic solar power forecasting using a hybrid Wavelet-PSO-SVM model based on SCADA and Meteorological information”, Renewable Energy, Vol. 118, pp. 357–367.

Fernandez-Jimenez, L.A., Muñoz-Jimenez, A., Falces, A., Mendoza-Villena, M., Garcia-Garrido, E., Lara-Santillan, P.M., Zorzano-Alba, E. and Zorzano-Santamaria, P.J. (2012), “Short-term power forecasting system for photovoltaic plants”, Renewable Energy, Vol. 44, pp. 311–317.

Gao, M., Li, J., Hong, F. and Long, D. (2019), “Day-ahead power forecasting in a large-scale photovoltaic plant based on weather classification using LSTM”, Energy, Vol. 187, p. 115838.

Gigoni, L., Betti, A., Crisostomi, E., Franco, A., Tucci, M., Bizzarri, F. and Mucci, D. (2018), “Day-Ahead Hourly Forecasting of Power Generation From Photovoltaic Plants”, IEEE Transactions on Sustainable Energy, Vol. 9 No. 2, pp. 831–842.

Giorgi, M.G. de, Congedo, P.M. and Malvoni, M. (2014), “Photovoltaic power forecasting using statistical methods: impact of weather data”, IET Science, Measurement & Technology, Vol. 8 No. 3, pp. 90–97.

Hanmin Sheng, J. Xiao, Y. Cheng, Qiang Ni and S. Wang (2018), “Short-Term Solar Power Forecasting Based on Weighted Gaussian Process Regression”, undefined.

Haque, A.U., Nehrir, M.H. and Mandal, P. (op. 2014), “Solar PV power generation forecast using a hybrid intelligent approach”, in Power and Energy Society General Meeting (PES), 2013 IEEE: Date 21-25 July 2013, Vancouver, BC, 7/21/2013 - 7/25/2013, IEEE, [S. l.], pp. 1–5.

Hossain, M.S. and Mahmood, H. (2020), “Short-Term Photovoltaic Power Forecasting Using an LSTM Neural Network and Synthetic Weather Forecast”, IEEE Access, Vol. 8, pp. 172524–172533.

Huang, C., Cao, L., Peng, N., Li, S., Zhang, J., Wang, L., Luo, X. and Wang, J.-H. (2018), “Day-Ahead Forecasting of Hourly Photovoltaic Power Based on Robust Multilayer Perception”, Sustainability, Vol. 10 No. 12, p. 4863.

Huang, Y.-C., Huang, C.-M., Chen, S.-J. and Yang, S.-P. (2020), “Optimization of Module Parameters for PV Power Estimation Using a Hybrid Algorithm”, IEEE Transactions on Sustainable Energy, Vol. 11 No. 4, pp. 2210–2219.

IEA (2020), “World Energy Outlook 2020 – Analysis - IEA”, available at: https://www.iea.org/reports/world-energy-outlook-2020 (accessed 18 March 2021).

Jesus, D.A.R. de, Mandal, P., Chakraborty, S. and Senjyu, T. (2019 - 2019), “Solar PV Power Prediction Using A New Approach Based on Hybrid Deep Neural Network”, in 2019 IEEE Power & Energy Society General Meeting (PESGM), Atlanta, GA, USA, 8/4/2019 - 8/8/2019, IEEE, pp. 1–5.

Kumar, A., Rizwan, M. and Nangia, U. (2020), “A Hybrid Intelligent Approach for Solar Photovoltaic Power Forecasting: Impact of Aerosol Data”, Arabian Journal for Science and Engineering, Vol. 45 No. 3, pp. 1715–1732.

Kumar, K.R. and Kalavathi, M.S. (2018), “Artificial intelligence based forecast models for predicting solar power generation”, Materials Today: Proceedings, Vol. 5 No. 1, pp. 796–802.

Kushwaha, V. and Pindoriya, N.M. (2017), “Very short-term solar PV generation forecast using SARIMA model: A case study”, in 2017 7th International Conference on Power Systems (ICPS), Pune, 12/21/2017 - 12/23/2017, IEEE, [Place of publication not identified], pp. 430–435.

Larson, D.P., Nonnenmacher, L. and Coimbra, C.F. (2016), “Day-ahead forecasting of solar power output from photovoltaic plants in the American Southwest”, Renewable Energy, Vol. 91, pp. 11–20.

Lee, D. and Kim, K. (2019), “Recurrent Neural Network-Based Hourly Prediction of Photovoltaic Power Output Using Meteorological Information”, Energies, Vol. 12 No. 2, p. 215.

Leva, S., Dolara, A., Grimaccia, F., Mussetta, M. and Ogliari, E. (2017), “Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power”, Mathematics and Computers in Simulation, Vol. 131, pp. 88–100.

Li, Z., Zang, C., Zeng, P., Yu, H. and Li, H. (2015 - 2015), “Day-ahead hourly photovoltaic generation forecasting using extreme learning machine”, in 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), Shenyang, China, 6/8/2015 - 6/12/2015, IEEE, pp. 779–783.

Liu, J., Fang, W., Zhang, X. and Yang, C. (2015), “An Improved Photovoltaic Power Forecasting Model With the Assistance of Aerosol Index Data”, IEEE Transactions on Sustainable Energy, Vol. 6 No. 2, pp. 434–442.

Liu, L., Zhan, M. and Bai, Y. (2019), “A recursive ensemble model for forecasting the power output of photovoltaic systems”, Solar Energy, Vol. 189, pp. 291–298.

Lu, H.J. and Chang, G.W. (2018), “A Hybrid Approach for Day-Ahead Forecast of PV Power Generation”, IFAC-PapersOnLine, Vol. 51 No. 28, pp. 634–638.

M. A. F. Lima, P. Carvalho, A. Braga, Luis M. Fernández Ramírez and Josileudo R. Leite (2018), “MLP Back Propagation Artificial Neural Network for Solar Resource Forecasting in Equatorial Areas”.

Madan Mohan Tripathi, Yash Pal and Harendra Kumar Yadav (2019), PSO tuned ANFIS model for short term photovoltaic power forecasting, Vol. 7.

Marquez, R. and Coimbra, C.F.M. (2013), “Proposed Metric for Evaluation of Solar Forecasting Models”, Journal of Solar Energy Engineering, Vol. 135 No. 1.

Massidda, L. and Marrocu, M. (2017), “Use of Multilinear Adaptive Regression Splines and numerical weather prediction to forecast the power output of a PV plant in Borkum, Germany”, Solar Energy, Vol. 146, pp. 141–149.

Massucco, S., Mosaico, G., Saviozzi, M. and Silvestro, F. (2019), “A Hybrid Technique for Day-Ahead PV Generation Forecasting Using Clear-Sky Models or Ensemble of Artificial Neural Networks According to a Decision Tree Approach”, Energies, Vol. 12 No. 7, p. 1298.

Mellit, A. and Kalogirou, S.A. (2008), “Artificial intelligence techniques for photovoltaic applications: A review”, Progress in energy and combustion science, Vol. 34 No. 5, pp. 574–632.

Mellit, A., Massi Pavan, A., Ogliari, E., Leva, S. and Lughi, V. (2020), “Advanced Methods for Photovoltaic Output Power Forecasting: A Review”, Applied Sciences, Vol. 10 No. 2, p. 487.

Mellit, A. and Pavan, A.M. (2010), “A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy”, Solar Energy, Vol. 84 No. 5, pp. 807–821.

Mishra, M., Byomakesha Dash, P., Nayak, J., Naik, B. and Kumar Swain, S. (2020), “Deep learning and wavelet transform integrated approach for short-term solar PV power prediction”, Measurement, Vol. 166, p. 108250.

Nespoli, A., Mussetta, M., Ogliari, E., Leva, S., Fernández-Ramírez, L. and García-Triviño, P. (2019), “Robust 24 Hours ahead Forecast in a Microgrid: A Real Case Study”, Electronics, Vol. 8 No. 12, p. 1434.

Nikodinoska, D., Käso, M. and Müsgens, F., “Solar and Wind Power Generation Forecasts using Elastic Net in Time-Varying Forecast Combinations”.

Ogliari, E., Dolara, A., Manzolini, G. and Leva, S. (2017), “Physical and hybrid methods comparison for the day ahead PV output power forecast”, Renewable Energy, Vol. 113, pp. 11–21.

Ogliari, E. and Nespoli, A. (2020), “Photovoltaic Plant Output Power Forecast by Means of Hybrid Artificial Neural Networks”, in Mellit, A. and Benghanem, M. (Eds.), A Practical Guide for Advanced Methods in Solar Photovoltaic Systems, Advanced Structured Materials, Vol. 128, Springer International Publishing, Cham, pp. 203–222.

Pazikadin, A.R., Rifai, D., Ali, K., Malik, M.Z., Abdalla, A.N. and Faraj, M.A. (2020), “Solar irradiance measurement instrumentation and power solar generation forecasting based on Artificial Neural Networks (ANN): A review of five years research trend”, The Science of the total environment, Vol. 715, p. 136848.

Pedro, H.T. and Coimbra, C.F. (2012), “Assessment of forecasting techniques for solar power production with no exogenous inputs”, Solar Energy, Vol. 86 No. 7, pp. 2017–2028.

Perveen, G., Rizwan, M., Goel, N. and Anand, P. (2020), “Artificial neural network models for global solar energy and photovoltaic power forecasting over India”, Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, pp. 1–26.

Pierro, M., Bucci, F., Felice, M. de, Maggioni, E., Moser, D., Perotto, A., Spada, F. and Cornaro, C. (2016), “Multi-Model Ensemble for day ahead prediction of photovoltaic power generation”, Solar Energy, Vol. 134, pp. 132–146.

Rajagukguk, R.A., Ramadhan, R.A.A. and Lee, H.-J. (2020), “A Review on Deep Learning Models for Forecasting Time Series Data of Solar Irradiance and Photovoltaic Power”, Energies, Vol. 13 No. 24, p. 6623.

Rana, M. and Rahman, A. (2020), “Multiple steps ahead solar photovoltaic power forecasting based on univariate machine learning models and data re-sampling”, Sustainable Energy, Grids and Networks, Vol. 21, p. 100286.

Raza, M.Q., Mithulananthan, N., Li, J., Lee, K.Y. and Gooi, H.B. (2019), “An Ensemble Framework for Day-Ahead Forecast of PV Output Power in Smart Grids”, IEEE Transactions on Industrial Informatics, Vol. 15 No. 8, pp. 4624–4634.

Raza, M.Q., Nadarajah, M. and Ekanayake, C. (2016), “On recent advances in PV output power forecast”, Solar Energy, Vol. 136, pp. 125–144.

Sangrody, H., Zhou, N. and Zhang, Z. (2020), “Similarity-Based Models for Day-Ahead Solar PV Generation Forecasting”, IEEE Access, Vol. 8, pp. 104469–104478.

Semero, Y.K., Zhang, J. and Zheng, D. (2018), “PV power forecasting using an integrated GA-PSO-ANFIS approach and Gaussian process regression based feature selection strategy”, CSEE Journal of Power and Energy Systems, Vol. 4 No. 2, pp. 210–218.

Sobri, S., Koohi-Kamali, S. and Rahim, N.A. (2018), “Solar photovoltaic generation forecasting methods: A review”, Energy Conversion and Management, Vol. 156, pp. 459–497.

Tao, C., Shanxu, D. and Changsong, C. (2010), “Forecasting power output for grid-connected photovoltaic power system without using solar radiation measurement”, in 2010 2nd IEEE International Symposium on Power Electronics for Distributed Generation Systems: PEDG 2010 ; Hefei, China, 16-18 June 2010, Hefei, China, 6/16/2010 - 6/18/2010, IEEE, Piscataway, NJ, pp. 773–777.

Theocharides, S., Makrides, G., Livera, A., Theristis, M., Kaimakis, P. and Georghiou, G.E. (2020), “Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing”, Applied Energy, Vol. 268, p. 115023.

Vagropoulos, S.I., Chouliaras, G.I., Kardakos, E.G., Simoglou, C.K. and Bakirtzis, A.G. (2016 - 2016), “Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting”, in 2016 IEEE International Energy Conference (ENERGYCON), Leuven, Belgium, 4/4/2016 - 4/8/2016, IEEE, pp. 1–6.

VanDeventer, W., Jamei, E., Thirunavukkarasu, G.S., Seyedmahmoudian, M., Soon, T.K., Horan, B., Mekhilef, S. and Stojcevski, A. (2019), “Short-term PV power forecasting using hybrid GASVM technique”, Renewable Energy, Vol. 140, pp. 367–379.

Varanasi, J. and Tripathi, M.M. (2019), “K-means clustering based photo voltaic power forecasting using artificial neural network, particle swarm optimization and support vector regression”, Journal of Information and Optimization Sciences, Vol. 40 No. 2, pp. 309–328.

Wang, F., Xuan, Z., Zhen, Z., Li, K., Wang, T. and Shi, M. (2020a), “A day-ahead PV power forecasting method based on LSTM-RNN model and time correlation modification under partial daily pattern prediction framework”, Energy Conversion and Management, Vol. 212, p. 112766.

Wang, J., Qian, Z., Wang, J. and Pei, Y. (2020b), “Hour-Ahead Photovoltaic Power Forecasting Using an Analog Plus Neural Network Ensemble Method”, Energies, Vol. 13 No. 12, p. 3259.

Yadav, A.K. and Chandel, S.S. (2017), “Identification of relevant input variables for prediction of 1-minute time-step photovoltaic module power using Artificial Neural Network and Multiple Linear Regression Models”, Renewable and Sustainable Energy Reviews, Vol. 77, pp. 955–969.

Yadav, H.K., Pal, Y. and Tripathi, M.M. (2020), “Short-term PV power forecasting using empirical mode decomposition in integration with back-propagation neural network”, Journal of Information and Optimization Sciences, Vol. 41 No. 1, pp. 25–37.

Yang, D. and Dong, Z. (2018), “Operational photovoltaics power forecasting using seasonal time series ensemble”, Solar Energy, Vol. 166, pp. 529–541.

Yang, D., Kleissl, J., Gueymard, C.A., Pedro, H.T. and Coimbra, C.F. (2018), “History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining”, Solar Energy, Vol. 168, pp. 60–101.

Yang, H.-T., Huang, C.-M., Huang, Y.-C. and Pai, Y.-S. (2014), “A Weather-Based Hybrid Method for 1-Day Ahead Hourly Forecasting of PV Power Output”, IEEE Transactions on Sustainable Energy, Vol. 5 No. 3, pp. 917–926.

Yu, D., Choi, W., Kim, M. and Liu, L. (2020), “Forecasting Day-Ahead Hourly Photovoltaic Power Generation Using Convolutional Self-Attention Based Long Short-Term Memory”, Energies, Vol. 13 No. 15, p. 4017.

Zang, H., Cheng, L., Ding, T., Cheung, K.W., Wei, Z. and Sun, G. (2020), “Day-ahead photovoltaic power forecasting approach based on deep convolutional neural networks and meta learning”, International Journal of Electrical Power & Energy Systems, Vol. 118, p. 105790.

Appendix A: The list of papers for data extraction

This appendix presents the list of 66 papers from which we have extracted the data for the statistical analysis in this paper.

Table 5: The list of papers for data extraction

Authors (Year)	Paper
Acharya *et al.* (2020)	Day-Ahead Forecasting for Small-Scale Photovoltaic Power Based on Similar Day Detection with Selective Weather Variables
Almonacid *et al.* (2014)	A methodology based on dynamic artificial neural network for short-term forecasting of the power output of a PV generator
Asrari *et al.* (2017)	A Hybrid Algorithm for Short-Term Solar Power Prediction—Sunshine State Case Study
Baharin *et al.* (2016)	Short-term forecasting of solar photovoltaic output power for tropical climate using ground-based measurement data
Bouzerdoum *et al.* (2013)	A hybrid model (SARIMA–SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant
Chen *et al.* (2011)	Online 24-h solar power forecasting based on weather type classification using artificial neural network
Chen *et al.* (2020)	Hour-ahead photovoltaic power forecast using a hybrid GRALSTM model based on multivariate meteorological factors and historical power datasets
Chupong and Plangklang (2011)	Forecasting power output of PV grid connected system in Thailand without using solar radiation measurement
Da Liu and Sun (2019)	Random forest solar power forecast based on classification optimization
Da Silva Fonseca *et al.* (2012)	Use of support vector regression and numerically predicted cloudiness to forecast power output of a photovoltaic power plant in Kitakyushu, Japan
Dan A. Rosa De Jesus *et al.* (2019)	Data Fusion Based Hybrid Deep Neural Network Method for Solar PV Power Forecasting
Das *et al.* (2017)	SVR-Based Model to Forecast PV Power Generation under Different Weather Conditions
Ding *et al.* (2011)	An ANN-based Approach for Forecasting the Power Output of Photovoltaic System
Dokur (2020)	Swarm Decomposition Technique Based Hybrid Model for Very Short-Term Solar PV Power Generation Forecast
E. Lorenz *et al.* (2007)	Forecast Of Ensemble Power Production By Grid-Connected Pv Systems
Eseye *et al.* (2018)	Short-term photovoltaic solar power forecasting using a hybrid Wavelet-PSO-SVM model based on SCADA and Meteorological information
Fernandez-Jimenez *et al.* (2012)	Short-term power forecasting system for photovoltaic plants
Gao *et al.* (2019)	Day-ahead power forecasting in a large-scale photovoltaic plant based on weather classification using LSTM
Gigoni *et al.* (2018)	Day-Ahead Hourly Forecasting of Power Generation From Photovoltaic Plants
Giorgi *et al.* (2014)	Photovoltaic power forecasting using statistical methods: impact of weather data
Hanmin Sheng *et al.* (2018)	Short-Term Solar Power Forecasting Based on Weighted Gaussian Process Regression
Haque *et al.* (op. 2014)	Solar PV Power Generation Forecast Using a Hybrid Intelligent Approach
Hossain and Mahmood (2020)	Short-Term Photovoltaic Power Forecasting Using an LSTM Neural Network and Synthetic Weather Forecast
Huang *et al.* (2018)	Day-Ahead Forecasting of Hourly Photovoltaic Power Based on Robust Multilayer Perception
Huang *et al.* (2020)	Optimization of Module Parameters for PV Power Estimation Using a Hybrid Algorithm
Jesus *et al.* (2019 - 2019)	Solar PV Power Prediction Using A New Approach Based on Hybrid Deep Neural Network
Kumar and Kalavathi (2018)	Artificial intelligence based forecast models for predicting solar power generation
Kumar *et al.* (2020)	A Hybrid Intelligent Approach for Solar Photovoltaic Power Forecasting: Impact of Aerosol Data
Kushwaha and Pindoriya (2017)	Very Short-Term Solar PV Generation Forecast Using SARIMA Model: A Case Study
Larson *et al.* (2016)	Day-ahead forecasting of solar power output from photovoltaic plants in the American Southwest
Lee and Kim (2019)	Recurrent Neural Network-Based Hourly Prediction of Photovoltaic Power Output Using Meteorological Information
Leva *et al.* (2017)	Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power

Li *et al.* (2015 - 2015)	Day-ahead Hourly Photovoltaic Generation Forecasting using Extreme Learning Machine
Liu *et al.* (2015)	An Improved Photovoltaic Power Forecasting Model With the Assistance of Aerosol Index Data
Liu *et al.* (2019)	A recursive ensemble model for forecasting the power output of photovoltaic systems
Lu and Chang (2018)	A Hybrid Approach for Day-Ahead Forecast of PV Power Generation
M. A. F. Lima *et al.* (2018)	MLP Back Propagation Artificial Neural Network for Solar Resource Forecasting in Equatorial Areas
Madan Mohan Tripathi *et al.* (2019)	PSO Tuned ANFIS Model for Short Term Photovoltaic Power Forecasting
Massidda and Marrocu (2017)	Use of Multilinear Adaptive Regression Splines and numerical weather prediction to forecast the power output of a PV plant in Borkum, Germany
Massucco *et al.* (2019)	A Hybrid Technique for Day-Ahead PV Generation Forecasting Using Clear-Sky Models or Ensemble of Artificial Neural Networks According to a Decision Tree Approach
Mellit and Pavan (2010)	A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy
Mishra *et al.* (2020)	Deep learning and wavelet transform integrated approach for short-term solar PV power prediction
Nespoli *et al.* (2019)	Robust 24 Hours ahead Forecast in a Microgrid: A Real Case Study
Nikodinoska *et al.* (*forthcoming*)	Solar and Wind Power Generation Forecasts using Elastic Net in Time-Varying Forecast Combinations
Ogliari and Nespoli (2020)	Photovoltaic Plant Output Power Forecast by Means of Hybrid Artificial Neural Networks
Ogliari *et al.* (2017)	Physical and hybrid methods comparison for the day ahead PV output power forecast
Pedro and Coimbra (2012)	Assessment of forecasting techniques for solar power production with no exogenous inputs
Perveen *et al.* (2020)	Artificial neural network models for global solar energy and photovoltaic power forecasting over India
Pierro *et al.* (2016)	Multi-Model Ensemble for day ahead prediction of photovoltaic power generation
Rana and Rahman (2020)	Multiple steps ahead solar photovoltaic power forecasting based on univariate machine learning models and data re-sampling
Raza *et al.* (2019)	An Ensemble Framework for Day-Ahead Forecast of PV Output Power in Smart Grids
Sangrody *et al.* (2020)	Similarity-Based Models for Day-Ahead Solar PV Generation Forecasting
Semero *et al.* (2018)	PV Power Forecasting Using an Integrated GA-PSO-ANFIS Approach and Gaussian Process Regression Based Feature Selection Strategy
Tao *et al.* (2010)	Forecasting Power Output for Grid-connected Photovoltaic Power System without using Solar Radiation Measurement
Theocharides *et al.* (2020)	Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing
Vagropoulos *et al.* (2016 - 2016)	Comparison of SARIMAX, SARIMA, Modified SARIMA and ANN-based Models for Short-Term PV Generation Forecasting
VanDeventer *et al.* (2019)	Short-term PV power forecasting using hybrid GASVM technique
Varanasi and Tripathi (2019)	K-means clustering based photo voltaic power forecasting using artificial neural network, particle swarm optimization and support vector regression
Wang *et al.* (2020a)	A day-ahead PV power forecasting method based on LSTM-RNN model and time correlation modification under partial daily pattern prediction framework
Wang *et al.* (2020b)	Hour-Ahead Photovoltaic Power Forecasting Using an Analog Plus Neural Network Ensemble Method
Yadav and Chandel (2017)	Identification of relevant input variables for prediction of 1-minute timestep photovoltaic module power using Artificial Neural Network and Multiple Linear Regression Models
Yadav *et al.* (2020)	Short-term PV power forecasting using empirical mode decomposition in integration with backpropagation neural network
Yang and Dong (2018)	Operational photovoltaics power forecasting using seasonal time series ensemble
Yang *et al.* (2014)	A Weather-Based Hybrid Method for 1-Day Ahead Hourly Forecasting of PV Power Output
Yu *et al.* (2020)	Forecasting Day-Ahead Hourly Photovoltaic Power Generation Using Convolutional Self-Attention Based Long Short-Term Memory
Zang *et al.* (2020)	Day-ahead photovoltaic power forecasting approach based on deep convolutional neural networks and meta learning

Appendix B: Model Classification

In this paper, we follow the model classification approach that is suggested by many other scholars (e.g., Rajagukguk *et al.* (2020), Antonanzas *et al.* (2016), and Sobri *et al.* (2018)), dividing models into 3 categories: (1) physical models, (2) statistical models, and (3) combination of models or "complex models".

First, physical models, also called PV performance, parametric, or "white box" method, use mathematical and physical mechanisms to predict PV power based on the information of many meteorological parameters. There are 3 main types of physical models including numerical weather prediction (NWP) model, sky imagery model, and satellite imaging model, with NWP being the most popular (Rajagukguk *et al.*, 2020).

Second, statistical models include all the models that use statistical data (usually the historical PV output data, possibly combined with meteorological variables) for their inputs and try to figure out the relationship of the data to forecast the time series of PV output. Under this category, we distinguish between persistence models, classical models including simple and advanced classical models, and ML models.

The persistence model, also known as the naïve or the elementary model, is the simplest form of the statistical model. It assumes that PV power output at time (t) the next day equals the PV output at the same time (t) of the previous day, which means the only input is the historical PV output data. For most of the cases, scholars claim that their proposed model outperform a range of other models, including the persistence. Because of the assumption that the value today equals tomorrow, the persistence model is not robust to the change in the weather conditions and only has the good performance on sunny days for very short-term forecasts.

Simple classical methodologies include mainly regression and autoregressive models (AR), and their extensions such as (non-linear) AR using exogenous variables (N-) ARX, (seasonal) AR moving integrated average (S-)ARIMA, and (S-)ARIMAX. The extension versions usually handle the non-stationary data better and therefore perform better than the basic AR models.

Advanced classical methodologies consist of the classical models that are combined with data processing and optimization techniques such as wavelet transformation, LOESS decomposition, gaussian regression, or exponential trend smoothing (ETS).

ML techniques are well-known for their better handling the complex non-linear relationship between multiple inputs and outputs and abilities of self-adaptation and inference, accompanied by more complexity and computational burden. The most popular ML models are ANN-based models, followed by SVM, random forest, and an increasing number of newly proposed models (Rajagukguk *et al.*, 2020).

Finally, the combination of different methods and techniques is the most advanced and complex methodology, including hybrid, ensemble, and hybrid-ensemble models. Hybrid method or also called "grey box" combines physical and statistical methods, with the outputs of one model being the inputs of the other models, and possibly together with multiple optimization techniques, while ensemble is more about combining forecast outputs from many individual models. Hybrid-ensemble is the combination of the two.

We focus our analysis on comparing the performance of the classical models (including both simple and advanced classical) and the state-of-the-art methodologies (including the ML, hybrid, ensemble, and hybrid-ensemble models).

Table 6 summarises the model classification and presents some examples of the models that are observed in the data set.

Table 6: PV output forecast models classification

<table>
<tr><th colspan="3">Model classification</th><th>Model</th></tr>
<tr><td colspan="3">Physical</td><td>NWP (E. Lorenz et al., 2007; Ogliari et al., 2017)</td></tr>
<tr><td rowspan="5">Statistical</td><td colspan="2">Persistence</td><td>Almost all papers in the list of Appendix A use persistence as among the benchmark models</td></tr>
<tr><td rowspan="2">Classical</td><td>Simple</td><td>ARIMA (Pedro and Coimbra, 2012; Tao et al., 2010), NARX (Tao et al., 2010), SARIMA (Vagropoulos et al., 2016 - 2016)</td></tr>
<tr><td>Advanced</td><td>ETS-based model (Zang et al., 2020), Gaussian-based regression (Da Liu and Sun, 2019), MARS (Massidda and Marrocu, 2017), Theta model (Yang and Dong, 2018)</td></tr>
<tr><td rowspan="4">State-of-the-art</td><td>ML</td><td>ANFIS (Kumar and Kalavathi, 2018), BPNN (Yadav et al., 2020), CFNN/ GRNN/RBFNN/ENN (Perveen et al., 2020), CNN/ RNN/FCNN (Jesus et al., 2019 - 2019), CSLSTM (Yu et al., 2020), DNN (Lee and Kim, 2019), FFNN (Dokur, 2020), kNN (Gigoni et al., 2018), LSTM (Chen et al., 2020), MLP (Liu et al., 2019), SVM (Wang et al., 2020a)</td></tr>
<tr><td rowspan="3">Complex</td><td>Hybrid</td><td>ARMAX-ANFIS-LSTM-FCN (Dan A. Rosa De Jesus et al., 2019), BPNN-TCM (Wang et al., 2020a), BP-SFLA-ANN (Asrari et al., 2017), GA-PSO-ANFIS (Semero et al., 2018), GRA-LSTM (Chen et al., 2020), K-means-ANN-PSO (Varanasi and Tripathi, 2019), WT-FNN-PSO (Raza et al., 2019), WT-LSTM-dropout network (Mishra et al., 2020)</td></tr>
<tr><td>Ensemble</td><td>Ensemble-Avg/Ensemble-OLS/Ensemble-LAD/Ensemble-CLS/Ensemble-lasso (Yang and Dong, 2018), Ensemble SARIMA(X) (Vagropoulos et al., 2016 - 2016)</td></tr>
<tr><td>Hybrid-ensemble</td><td>CSM-ANN-Ensemble-CART (Massucco et al., 2019), Ensemble-Avg-post, Ensemble-CLS-post/Ensemble-LAD-post/Ensemble-lasso-post/Ensemble-OLS-post (Yang and Dong, 2018)</td></tr>
</table>

Appendix C: Error metric formula

This appendix shows the formulas of the error metrics whose data we extracted from the literatures, including the NRMSE, NMAE, and MAPE. The formulas presented below are observed as the standard formulas of these error metrics as provided by the individual papers on PV output forecasts listed in Table 5.

$$NRMSE\_avg(\%) = \frac{\sqrt{\frac{1}{N}\sum_{1=1}^{N}(\widehat{p_i} - p_i)^2}}{\bar{p}} * 100 \quad (3)$$

$$NRMSE\_installed\ (\%) = \frac{\sqrt{\frac{1}{N}\sum_{1=1}^{N}(\widehat{p_i} - p_i)^2}}{p_{installed/peak}} * 100 \quad (4)$$

$$NRMSE\_norm\ (\%) = \sqrt{\frac{1}{N}\sum_{1=1}^{N}(\widehat{n_i} - n_i)^2} * 100 \quad (5)$$

$$NMAE\_avg(\%) = \frac{\frac{1}{N}\sum_{i=1}^{N}|\widehat{p_i} - p_i|}{\bar{p}} * 100 \quad (6)$$

$$NMAE\_installed(\%) = \frac{\frac{1}{N}\sum_{i=1}^{N}|\widehat{p_i} - p_i|}{p_{installed/peak}} * 100 \quad (7)$$

$$NMAE\_norm(\%) = \frac{1}{N}\sum_{i=1}^{N}|\widehat{n_i} - n_i| * 100 \quad (8)$$

$$MAPE\_avg(\%) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\widehat{p_i} - p_i}{p_i}\right| * 100 \quad (9)$$

$$MAPE\_installed(\%) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\widehat{p_i} - p_i}{p_{installed/peak}}\right| * 100 \quad (10)$$

$$MAPE\_norm(\%) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\widehat{n_i} - n_i}{n_i}\right| * 100 \quad (11)$$

where N is the total number of forecast points in the forecasting period, i represents the time step, $\widehat{p_i}$ and p_i represent the forecast and actual values of PV output at the time step i, $\bar{p}$ stands for the mean value of PV output, $p_{installed/peak}$ indicates the installed capacity of the PV plant or the peak power achieved by the plant, and $\widehat{n_i}$ and n_i are the normalized forecast and actual PV output calculated based on the normalized input data at the time step i.